2 Charakterizace jedné náhodné veličiny

Cílem tématu je seznámení se základními typy pravděpodobnostních rozdělení a jejich charakteristikami. Dále se základními statistikami a jejich užitím u veličin různých typů, rozdělením četností (tabulky, grafy), s pojmy výběrová distribuční funkce, bodové a intervalové odhady parametrů, testy o parametrech základních rozdělení.

2.1 Pravděpodobnost a náhodná veličina

Jestliže jsme data výběrového souboru získali na základě dobře navrženého experimentu, lze na jejich základě provádět *zobecňující úsudky* o chování základního souboru (populace). Konkrétně jde o úsudky o proměnných a jejich parametrech, které populaci popisují. Takové metody vycházejí z počtů pravděpodobnosti a vyžadují data získaná náhodným výběrem nebo znáhodněným experimentem. Pokud proces sběru dat nelze opakovat, tedy pokud si nedokážeme představit, že bychom z populace mohli vybrat jiný náhodný výběr, statistické usuzování nemá smysl.

Statistika se často zabývá událostmi, jejichž průběh a výsledek není definovaný přesnými, předem známými pravidly. Řada událostí je ovlivněna vlivy, které není možné podchytit nebo jsou pro nás tyto vlivy zatím neznámé a nepopsatelné. Významnou roli při vývoji řady událostí hraje *náhoda*.

Sledujeme-li výsledek nějaké události, který nastává opakovaně za určitých stejně nastavených podmínek, nazýváme každé takové pozorování *pokusem*. Může se jednat o děj probíhající v přírodě, ve výrobě, ve společnosti, v ekonomice či v laboratoři. Statisticky chápaný pojem pokus nemusí znamenat to samé co skutečný experiment, který je řízený pozorovatelem.

Poznámka: Při pokusech sledujeme chování reality při určitých podmínkách. V laboratoři pro každý pokus nastavíme snadno předem zvolenou teplotu. V přírodě se musíme při „pokusech“ spokojit s teplotou či teplotami, které se v průběhu pozorování právě vyskytují. Rovněž při ekonomických pokusech si nemůžeme libovolně nastavovat podmínky – inflaci, produktivitu práce, úrokovou míru. Nejedná se tudíž o pokusy ve smyslu experimentů, ale o šetření. Děje se opakují bez toho, aby výzkumník mohl nastavovat jejich podmínky. Může je pouze sledovat a vyhodnocovat.

Existují dva základní typy chování pozorované reality při pokusu:

1. Při stejném souboru vstupních podmínek se dostaví vždy stejný výsledek – *deterministický pokus* (např. hozená hrací kostka padá vždy k zemi).
2. Při stejném souboru vstupních podmínek se dostaví jeden z možných výsledků pokusu – *náhodný pokus*. Existuje určitá konečná či nekonečná množina možných výsledků pro daný náhodný pokus a jeden z nich náhodně nastane. Výsledek závisí nejen na nastavených podmínkách, jako u deterministického pokusu, ale i na náhodě (např. při házení kostky padne vždy 1 až 6 ok).

Náhodný jev je jeden z možných výsledků náhodného pokusu. Náhodný pokus může končit různými výsledky. Očekáváme, že se dostaví jeden z několika možných (alespoň dvou) náhodných jevů. Při dlouhodobějším - *hromadném sledování*, lze zjistit, že náhodné jevy nejsou zcela náhodné a tudíž bez jakýchkoliv zákonitostí. Naopak zjistíme, že se řídí *pravděpodobnostními zákony*. Podle těchto zákonů bohužel nelze určit, jaký konkrétní náhodný jev nastane, ale lze s jejich pomocí stanovit jaká je *pravděpodobnost*, že daný náhodný jev nastane. Lze také určit, kolikrát se při větším počtu pokusů bude přibližně daný náhodný jev opakovat.

Běžně využíváme pojmy pravděpodobný nebo nepravděpodobný, abychom kvalitativně vyjádřili jistotu určitého jevu. Pojem *pravděpodobnost* se postupem času vyvíjel až do podoby čistě matematicky definovaného pojmu. Pravděpodobnost se tak vyjadřuje číslem. Uveďme, že podle *matematické definice* mj. platí, že pravděpodobnost *jistého je* 1 (lze uvádět i jako 100%) a že pravděpodobnost náhodného jevu je nezáporné číslo. Pro další výklad je nejpodstatnější s*tatistická definice* (von Misessova).

Provedli jsme *n*-krát náhodný pokus, přičemž konkrétní náhodný jev (označme jej třeba A) nastal *m*-krát. Relativní četnost jevu A (tj. poměr *m*/*n*) se přibližuje (konverguje) k pravděpodobnosti tohoto jevu pro velký počet náhodných pokusů. Matematicky lze definici zapsat:

$$\mathrm{P(A)} = \lim_{n\to\infty} \frac{m}{n}.$$

Podle statistické definice je relativní četnost jevu odhadem pravděpodobnosti, že tento jev nastane. Relativní četnost určená na základě pozorování opakujících se náhodných jevů kolísá kolem hodnoty pravděpodobnosti a přibližuje se k ní s rostoucím počtem pokusů *n*. Proto je ve vzorci limita pro veliký počet pokusů $n \to \infty$. Toto přibližování má ovšem pouze pravděpodobnostní charakter, tj. s každým dalším pokusem se relativní četnost nemusí nutně více přiblížit k pravděpodobnosti. Nedostatkem definice je, že dle ní je možné pravděpodobnost určit jen na základě pokusu. Na druhé straně nám dává návod, jak lze opakovaným pozorováním odhadnout pravděpodobnost náhodného jevu.

Často používáme pojem *šance*. Jde o poměr ve prospěch nějakého náhodného jevu (A) počítaný jako:

$$\frac{\text{počet výskytů jevu A}}{\text{počet případů, kdy jev A nenastal}}.$$

Náhodná veličina je statistická veličina, která za stejných podmínek měření či pozorování může nabývat působením náhodných vlivů různých hodnot. Náhodná veličina je nástroj pro kvantitativní popis náhodných jevů.

V množině všech náhodných jevů, které mohou v rámci náhodného pokusu nastat (v elementárním prostoru náhodných jevů), se mohou vyskytovat jevy, které nastávají častěji a jejichž pravděpodobnost je vyšší ale také jevy méně časté s pravděpodobností nižší. Ve

statistice hovoříme o *rozdělení pravděpodobnosti náhodné veličiny* nebo zkráceně rozdělení pravděpodobnosti.

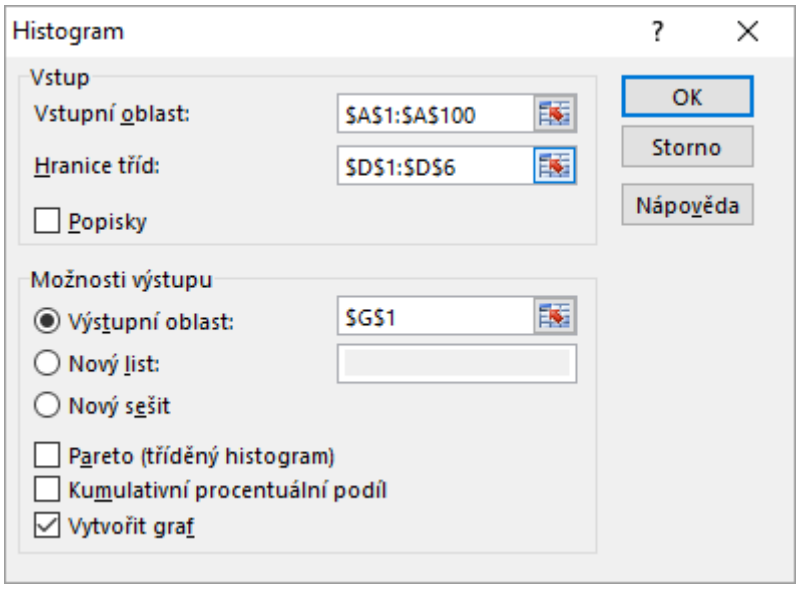
Sledováním nebo měřením náhodné veličiny lze stanovit rozdělení relativních četností naměřených hodnot (v kolika procentech případů se sledovaná hodnota skutečně objevila). Jde o *empirické rozdělení četností*. V určitých případech dokážeme spočítat i očekávané četnosti, s nimiž se hodnoty mají vyskytovat, pak jde o *teoretické rozdělení četností*. Rozdíl mezi oběma rozděleními je patrný ze statistické definice pravděpodobnosti. Hodnota relativní četnosti se blíží pravděpodobnosti, přičemž v podstatě se s rostoucím počtem pozorování rozdíl zmenšuje.

Příklad: Provedeme 100x náhodný pokus hod šestistěnnou kostkou a zapíšeme počty padlých ok. Na kostce může padnout jedno až šest ok. Hodnoty 1, 2, 3, 4, 5 a 6 tak tvoří prostor elementárních jevů (jiný počet ok padnout nemůže) a jsou náhodnou veličinou. Teoretické četnosti jsou pro všechny hodnoty náhodné veličiny stejné 1/6 tj. 0,167. Výsledky jsou uvedeny v tabulce. S rostoucím počtem pokusů se empirické četnosti relativní budou blížit četnostem teoretickým.

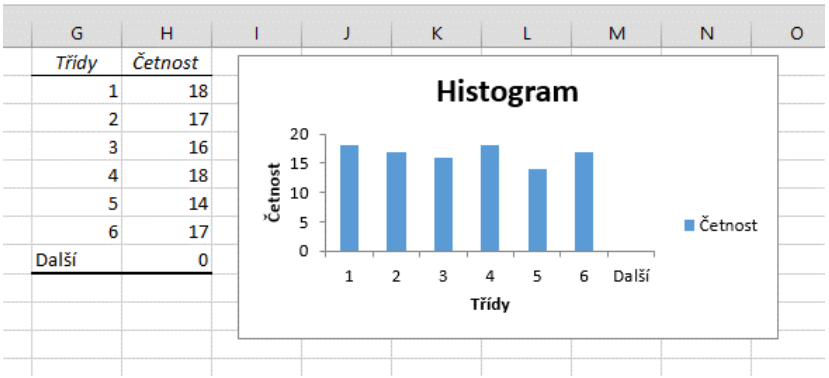
Náhodná veličina	**Empirická četnost**	**Empirická četnost relativní**	**Teoretická četnost**
1	18	0,18	0,167
2	17	0,17	0,167
3	16	0,16	0,167
4	18	0,18	0,167
5	14	0,14	0,167
6	17	0,17	0,167

Tab. Porovnání empirických a teoretických četností

V programu MS Excel lze z datového souboru vytvořit výše uvedenou tabulkou v nabídce Data → Analýza dat → Histogram. Do pole Vstupní oblast zadáme odkaz na oblast buněk, která obsahuje vstupní čísla, a do pole Hranice tříd zadáme odkaz na oblast buněk, která obsahuje čísla tříd (v našem případě jde o buňky, kde jsou postupně vypsány hodnoty 1 až 6). Do pole Výstupní oblast zadáme buňku, kam se má tabulka vypsat. Lze zobrazit i graf četností – histogram zaškrtnutím volby Vytvořit graf. Pozor nástroj počítá pouze absolutní četnosti. Relativní četnosti je nutno dopočítat dle jednoduchého vzorce *absolutní četnost/suma všech četností*.

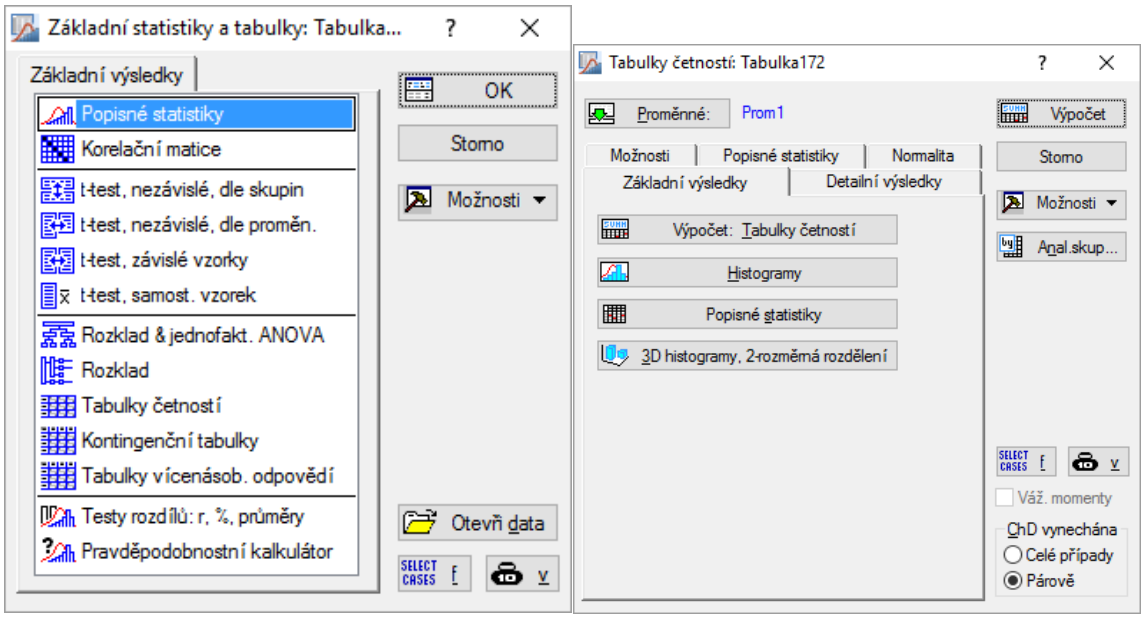


Obr. Zadávací okno nástroje Histogram programu MS Excel.

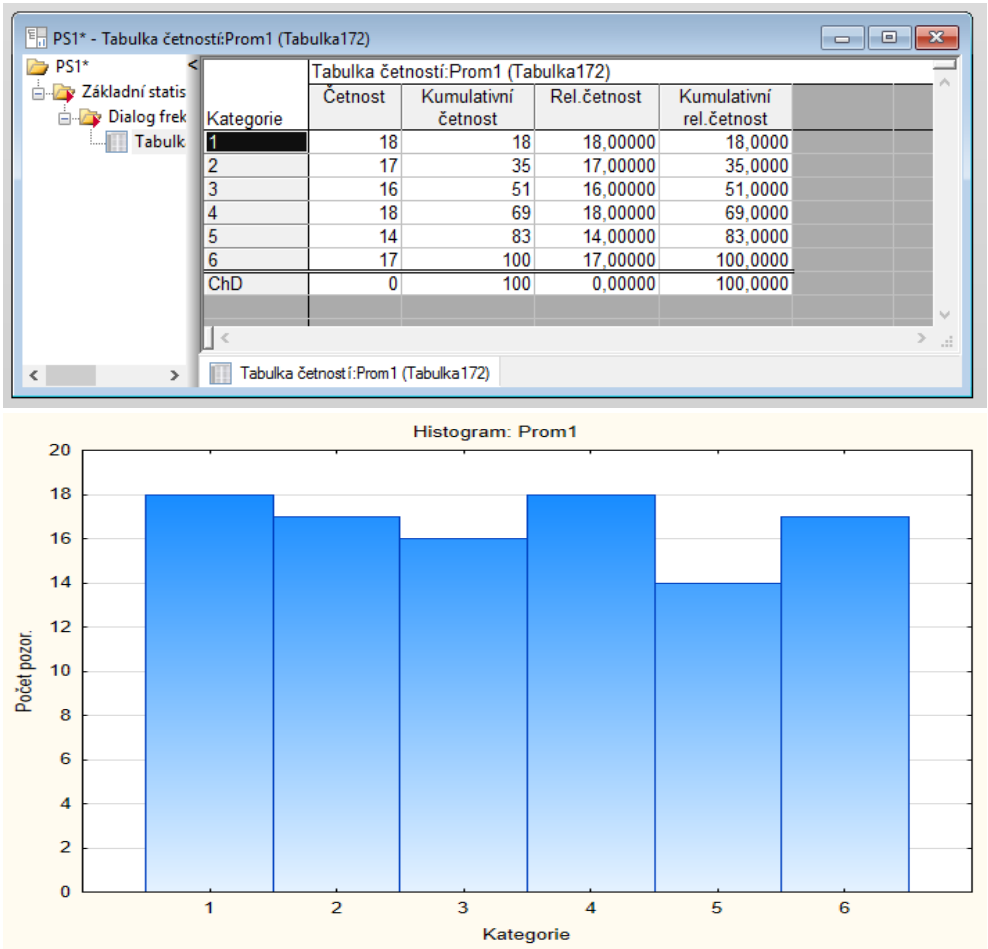


Obr. Výstup z nástroje Histogram programu MS Excel.

V programu Statistika lze z dat vytvořit výše uvedenou tabulkou v nabídce Statistiky → Základní statistiky/tabulky → Tabulky četností, kde zvolíme Výpočet: Tabulky četností. Pozor, na rozdíl od programu MS Excel jsou vypočteny i relativní četnosti a jsou uvedeny v procentech. Zobrazit lze i histogram v nabídce Histogramy.



Obr. Zadávací okna nástroje Tabulky četností programu Statistica.

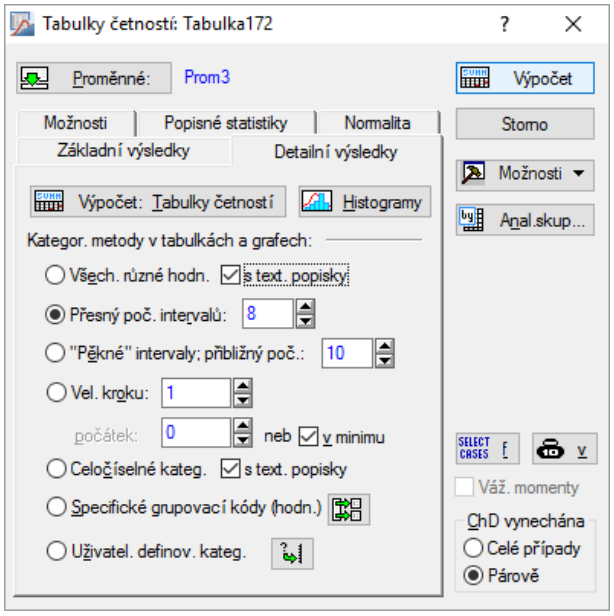


Obr. Výstupy z nástroje Tabulky četností programu Statistica.

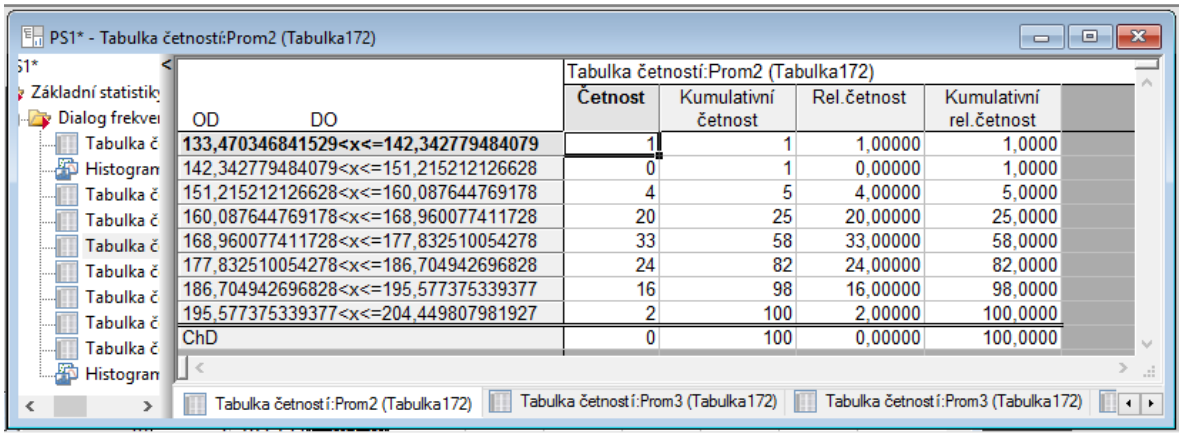
Příklad: Obdobně jako v kapitole 1 využijeme fiktivní situaci, kdy ze základního souboru výšek lidí (předpokládáme normální rozdělení se střední hodnotou 178 cm a směrodatnou odchylkou 10 cm) provedeme náhodný výběr 100 lidí. Tabulku četností lze v obou pro-

gramech vytvořit stejnými nástroji jako v předchozím příkladu. Není ale vhodné takovouto proměnnou uspořádávat (třídit) stejným způsobem. Mohla by nastat situace, kde by se v souboru každá hodnota vyskytovala pouze jednou a pak by všechny vypočtené četnosti měly hodnotu 1. Provedeme raději intervalové třídění, kdy si nejdříve stanovíme meze tříd, pro které četnosti necháme spočítat. Takovéto třídění je vhodné pro spojité náhodné veličiny. Počet tříd (počet řádků tabulky) a jejich meze se odvíjejí od počtu hodnot v souboru. Dle Sturgesova pravidla bude optimální počet tříd přibližně 1+3,3*log(*počet hodnot*). V našem případě tedy použijeme 8 tříd, protože 1+3,3*log(100)=7,6.

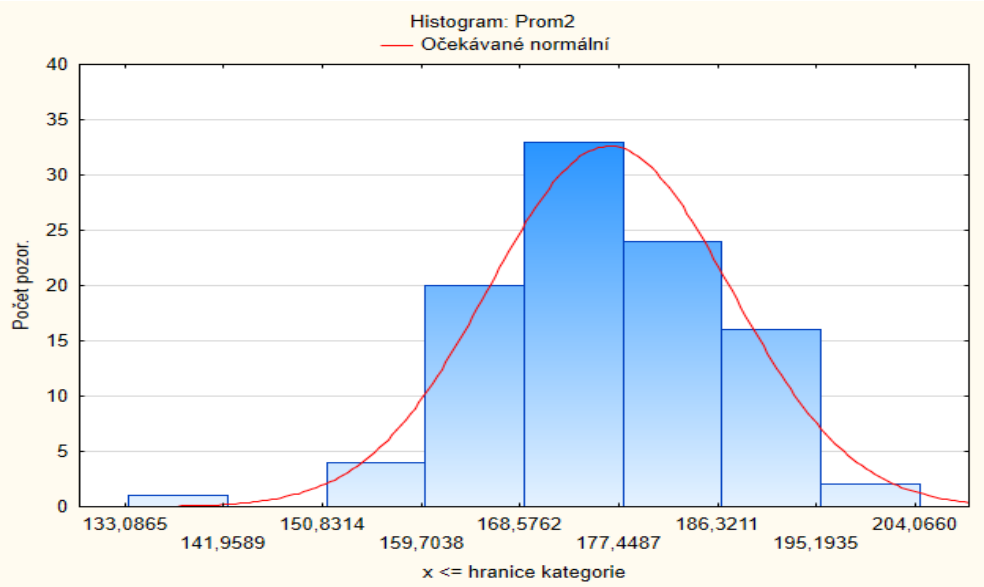
V programu Statistika opět využijeme nabídku Statistiky → Základní statistiky/tabulky → Tabulky četností, kde zvolíme kartu Detailní výsledky a vypočtený počet tříd zadáme do kolonky Přesný počet intervalů. Alternativně lze využít i kolonku Pěkné intervaly; přibližný počet. Poté potvrdíme tlačítkem Výpočet: Tabulky četností pro výpočet tabulky a/nebo tlačítkem Histogramy pro zobrazení grafu četností.



Obr. Zadávací okno nástroje Tabulky četností programu Statistica.



OD DO	Četnost	Kumulativní četnost	Rel.četnost	Kumulativní rel.četnost
133,470346841529<x<=142,342779484079	1	1	1,00000	1,0000
142,342779484079<x<=151,215212126628	0	1	0,00000	1,0000
151,215212126628<x<=160,087644769178	4	5	4,00000	5,0000
160,087644769178<x<=168,960077411728	20	25	20,00000	25,0000
168,960077411728<x<=177,832510054278	33	58	33,00000	58,0000
177,832510054278<x<=186,704942696828	24	82	24,00000	82,0000
186,704942696828<x<=195,577375339377	16	98	16,00000	98,0000
195,577375339377<x<=204,449807981927	2	100	2,00000	100,0000
ChD	0	100	0,00000	100,0000



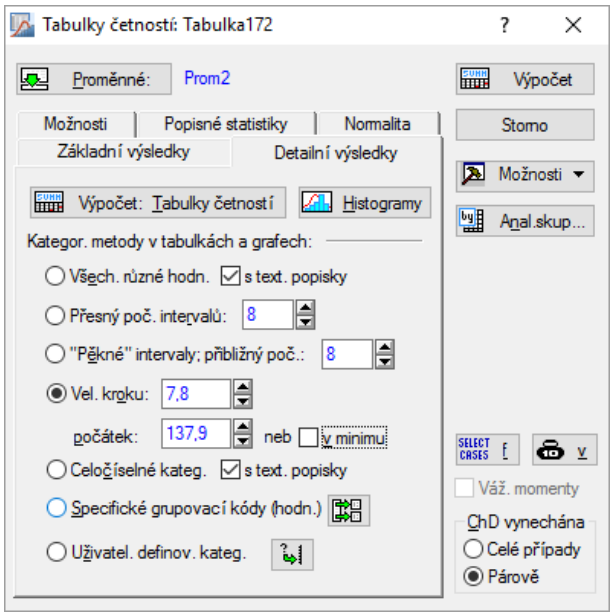
Obr. Výstupy z nástroje Tabulky četností programu Statistica.

Výsledky pokusu jsou uvedeny v následující tabulce. Teoretické četnosti byly vypočteny s využitím pravděpodobnostního počtu založeného na normálním rozdělení (viz další kapitola). S rostoucím počtem pokusů se empirické četnosti relativní budou blížit četnostem teoretickým.

Náhodná veličina	Empirická četnost	Empirická četnost relativní	Teoretická četnost
(133,47; 142,34>	1	0,01	0,00
(142,34; 151,22>	0	0,00	0,00
(151,22; 160,09>	4	0,04	0,03
(160,09; 168,96>	20	0,20	0,15
(168,96; 177,83>	33	0,33	0,31
(177,83; 186,7>	24	0,24	0,31

Tab. Intervalově uspořádaná tabulka četností.

Tvorba tabulky četností s intervalovým tříděním je v programu MS Excel náročnější v porovnání s programem Statistica. Nejprve je totiž nutné stanovit meze tříd. Vyjdeme opět ze Sturgesova pravidla s tím, že tabulku uspořádáme do 8 tříd. Zjistíme minimum (137,91 cm) a maximum (200,01 cm) souboru a rozdíl mezi nimi rozdělíme na 8 stejně širokých tříd. Výpočtem *(maximum-minimum)/počet tříd* zjistíme šířku každé z tříd (7,8 cm). Minimum souboru by se mělo nacházet v první třídě, tedy můžeme její meze stanovit jako (137,9; 145,7>. Pro zadání do programu stačí spočítat pouze horní meze tříd, tedy jde o hodnoty 145,7; 153,5; 161,3; 169,1; 176,9; 184,7; 192,5; 200,3 s tím, že každá hodnota bude v samostatné buňce. V nabídce Data → Analýza dat → Histogram zadáme do pole Vstupní oblast odkaz na oblast buněk, která obsahuje vstupní čísla, a do pole Hranice tříd zadáme odkaz na oblast buněk, která obsahuje stanovené horní meze tříd. Výsledná tabulka se z důvodu odlišně stanovených mezí liší od tabulky získané programem Statistika. Ve Statistice lze také použít výše uvedený postup a to tak, že vypočtenou šířku třídy (7,8) a dolní mez první třídy (137,9) zadáme v nabídce Statistiky → Základní statistiky/tabulky → Tabulky četností → Detailní výsledky do odpovídajících kolonek Vel. kroku.



Obr. Zadávací okno nástroje Tabulky četností programu Statistica.

Výše uvedené postupy tvorby tabulek četností a histogramů lze obecně provádět při průzkumové analýze dat. Základní pohled na histogram může být např. použit při posuzování normality rozdělení dat.

2.2 Základní typy pravděpodobnostních rozdělení a jejich charakteristiky

Můžeme si představit, že rozdělení pravděpodobností náhodné veličiny je matematický model, podle kterého jsou generovány hodnoty náhodné veličiny. Při náhodném pokusu pak určíme konkrétní rozdělení relativních četností. Rozdělení empirické se může od teoretického lišit jen v určitých mezích, v závislosti na počtu naměřených hodnot. Ve statistice se sleduje nebo určuje zákon rozdělení pravděpodobnosti náhodné veličiny. Při vyjádření tohoto zákona rozdělení je pro danou náhodnou veličinu třeba určit:

1. jakých hodnot může veličina nabývat a
2. s jakými pravděpodobnostmi se tyto hodnoty vyskytují.

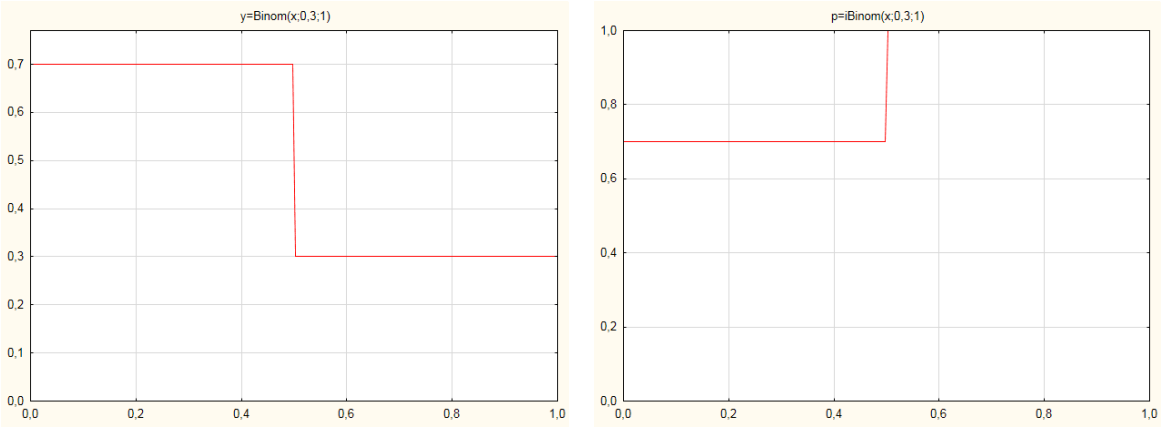
V případě nespojité náhodné veličiny přiřazujeme každé hodnotě pravděpodobnost výskytu. U spojité náhodné veličiny přiřazujeme pravděpodobnost výskytu určitému zvolenému intervalu hodnot. Sledované rozdělení pravděpodobnosti může být vyjádřeno matematickým vztahem nebo tabulkou všech možných hodnot s jejich příslušnými pravděpodobnostmi a nebo grafem. Zmíněných matematických vztahů může být více: pravděpodobnostní funkce, hustota pravděpodobnosti, funkce přežití. Praktický význam má zejména distribuční funkce, která má stejnou interpretaci u všech typů rozdělení. *Distribuční funkce* (kumulativní distribuční funkce) se značí F(x) a udává pravděpodobnost, že náhodná veličina nabude hodnoty menší nebo stejně velké jako je zvolená hodnota x. Funkce může na-

bývat pouze hodnot od 0 do 1 (včetně) a má neklesající průběh. V hodnotě blížící se -∞ je její hodnota 0 a v hodnotě blížící se +∞ je její hodnota 1. Funkce nemusí být spojitá.

Také rozdělení pravděpodobností náhodné veličiny mají své charakteristické vlastnosti: polohu, variabilitu, tvar. Jejich výpočty jsou specifické dle typu rozdělení a budou uvedeny dále v textu.

2.2.1 Alternativní rozdělení A(p)

Pokud náhodná veličina nabývá pouze dvou hodnot (0 a 1) - jde tedy o diskrétní rozdělení, hovoříme o alternativním rozdělení. Alternativním rozdělením se řídí například hod mincí (padne panna nebo orel), odpověď na otázku (odpovím správně nebo ne?), narození dítěte (narodí se holka nebo kluk?). Tedy pokusy se dvěma možnými výsledky. Pravděpodobnost výskytu hodnoty 1 je p a pravděpodobnost výskytu hodnoty 0 je q. Pro rozdělení platí $q = 1 - p$. Výskyt jiné hodnoty než 0 nebo 1 je jev nemožný. Alternativní rozdělení má tak jediný parametr - pravděpodobnost p. Pravděpodobnost (pravděpodobnostní funkci) lze zapsat rovnicí: $P(X = x) = p^x(1-p)^{(1-x)}$, kde x může být buď 1 nebo 0. Hodnoty pravděpodobnostní a distribuční funkce lze spočítat pomocí programů MS Excel nebo Statistica (viz přiložená tabulka na konci kapitoly). Střední hodnota rozdělení je odvozena z parametru p jako $E(X) = p$ a rozptyl rozdělení je $D(X) = p(1-p)$.



Obr. Pravděpodobnostní a distribuční funkce alternativního rozdělení s parametrem p = 0,3.

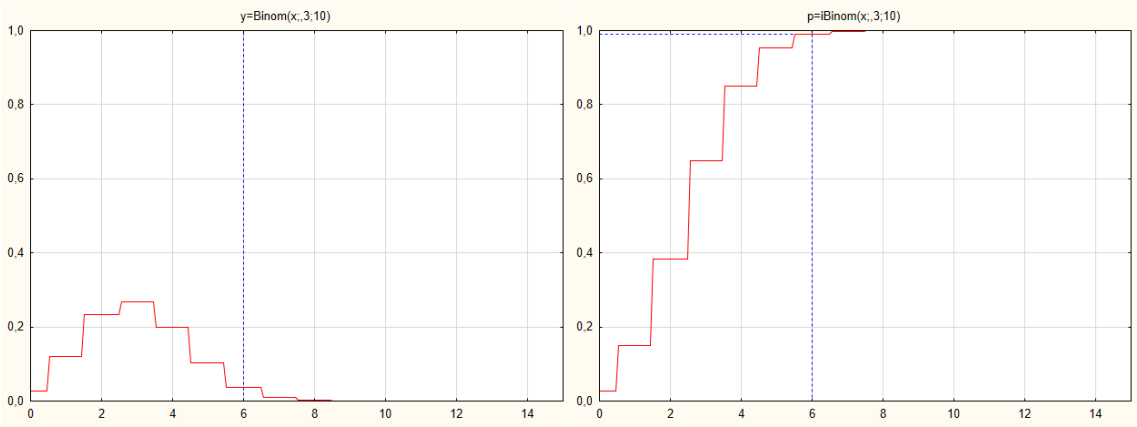
2.2.2 Binomické rozdělení Bi(p;n)

Binomické rozdělení se využívá k popisu náhodné veličiny udávající počet výskytů náhodného jevu (počet příznivých výsledků, úspěchů - jde tedy o diskrétní rozdělení), pokud náhodný pokus opakujeme vícekrát (n-krát) a pravděpodobnost p výskytu tohoto jevu je při každém opakování konstantní. Pravděpodobnost výskytu opačného jevu q musí pak být v každém pokusu také stále stejná ($q = 1 - p$). Jde tedy o rozšíření alternativního rozdělení na víc pokusů. Výsledek následného pokusu nesmí být ovlivněn výsledkem pokusu předchozího, tedy opakování musejí být vzájemně nezávislá. Typickým příkladem je tzv. *náhodný výběr s vracením prvků.*

Veličina X tedy představuje počet úspěchů a nabývá hodnot $x_i = 0, 1, 2$ až n. Pravděpodobnosti pro uvedené hodnoty x_i se počítají podle pravděpodobnostní funkce:

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}.$$

Hodnoty pravděpodobnostní a distribuční funkce lze spočítat pomocí programů MS Excel nebo Statistica (viz přiložená tabulka na konci kapitoly). Pro střední hodnotu binomického rozdělení platí $E(X) = np$ a pro rozptyl platí $D(X) = np(1\text{-}p)$.



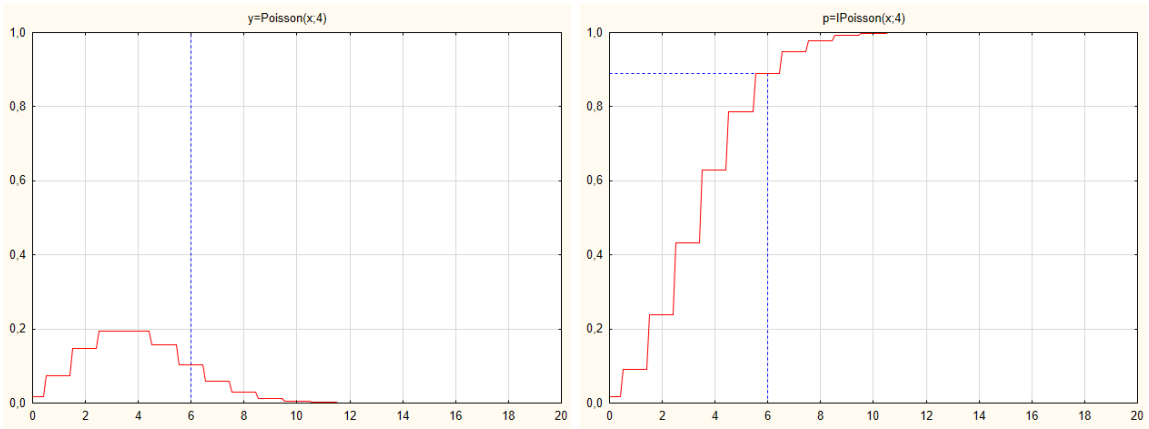
Obr. Pravděpodobnostní a distribuční funkce binomického rozdělení s parametry $n = 10$ a $p = 0{,}3$.

2.2.3 Poissonovo rozdělení Po(λ)

Poissonovo rozdělení se používá při sledování počtu událostí (jde tedy o diskrétní rozdělení) v jednotce času, plochy nebo objemu za podmínky, že se pravděpodobnost výskytu jevu nemění. Veličina může nabývat pouze nezáporných celých čísel. Pravděpodobnost lze spočítat pomocí funkce:

$$P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!}.$$

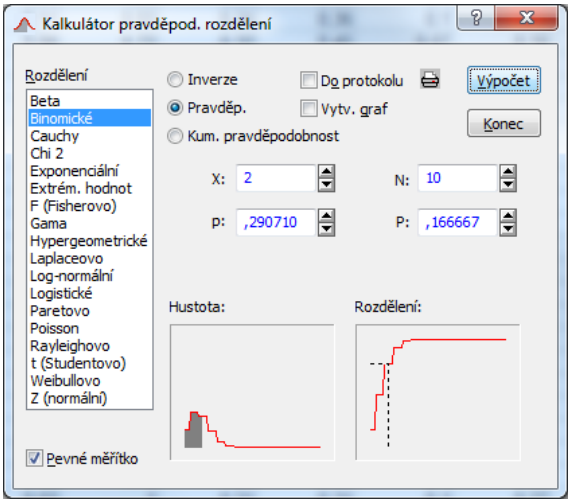
Poissonovo rozdělení má jediný parametr λ. Při praktickém použití určujeme parametr λ na základě toho, že představuje střední hodnotu náhodné veličiny. Odhadujeme ji průměrným počtem výskytů daného jevu ve vymezeném časovém intervalu nebo v jednotce prostoru. Hodnoty pravděpodobnostní a distribuční funkce lze spočítat pomocí programů MS Excel nebo Statistica (viz přiložená tabulka na konci kapitoly). Střední hodnota náhodné veličiny a rozptyl nabývají hodnoty parametru rozdělení $E(X) = \lambda$ a $D(X) = \lambda$.



Obr. Pravděpodobnostní a distribuční funkce Poissonova rozdělení s parametrem $\lambda = 4$.

Software Statistica počítá pravděpodobnosti v nástroji Kalkulátor pravděpod. rozdělení (Statistiky → Pravděpodobnostní kalkulátor → Rozdělení). U diskrétních rozdělení lze mj. počítat hodnotu distribuční funkce (značeno jako Kum. pravděpodobnost) a pravděpodobnostní funkci (značeno jako Pravděp.). U spojitých rozdělení se počítá distribuční funkce tak, že není zaškrtnuto žádné z nabízených políček.

Příklad: Výpočet pravděpodobnosti z binomického rozdělení s parametry: počet opakování 10 (na obrázku značeno N:), pravděpodobnost úspěchu při jednom pokusu 0,166667 (na obrázku P:). Počet úspěšných pokusů 2 (na obrázku X:). Pravděpodobnost je 0,290710 (na obrázku p:). Graf pravděpodobnosti je nazván jako Hustota, graf distribuční funkce je nazván jako Rozdělení.



Obr. Zadávací okno nástroje Tabulky četností programu Statistica.

Přehled funkcí, které jsou pro výpočet pravděpodobnosti k dispozici v Excelu, je v tabulce 2. Pro úplnost jsou do stejné tabulky vloženy i funkce programu Statistica.

Rozdělení	Parametry	MS Excel	Statistica
Alternativní A(p)	p – pravděpodobnost úspěchu	=*BINOMDIST*(1;1;p; 1[1]) [1]pro výpočet pravděpodobnostní funkce se zadá parametr 0)	Rozdělení: Binomické, kde N bude vždy 1, P je parametr p a X může být 0 nebo 1.
Binomické Bi(n;p)	p – pravděpodobnost úspěchu při jednom pokusu n – počet pokusů	=*BINOMDIST*(počet úspěšných pokusů x;n; p; 1[1]) [1]pro výpočet pravděpodobnostní funkce se zadá parametr 0)	Rozdělení: Binomické, kde N je parametr n, P je parametr p a X je libovolné celé nezáporné číslo.
Poissonovo Po(λ)	λ – průměrný počet událostí na jednotku času nebo prostoru	=*POISSON* (počet úspěšných událostí x; λ; 1[1]) [1]pro výpočet pravděpodobnostní funkce se zadá parametr 0)	Rozdělení: Poissonovo, kde Lambda je parametr λ a X je libovolné celé nezáporné číslo.
Normální $N(\mu;\sigma^2)$	μ – střední hodnota σ – směrodatná odchylka (odmocnina z rozptylu)	=*NORMDIST* (x;μ;σ;1[1]) [1]pro výpočet hustotní funkce se zadá parametr 0	Rozdělení: Z (normální), kde průměr je parametr μ, SmOdch je parametr σ a X je libovolné číslo.
Normované normální $N(0;1)$	μ = 0 střední hodnota σ = 1 směrodatná odchylka (odmocnina z rozptylu)	= *NORMSDIST*(x)	Rozdělení: Z (normální), kde průměr je 0, SmOdch je 0 a X je libovolné číslo.
Chí-kvadrát $\chi^2(v)$	v – počet stupňů volnosti	= *CHIDIST*(x;v)	Rozdělení: Chi 2, kde sv. je parametr v a chi 2 je libovolné nezáporné číslo.
Studentovo $t(v)$	v – počet stupňů volnosti	= *TDIST*(x;v)	Rozdělení: t (Studentovo), kde sv. je parametr v a t je libovolné číslo.
F rozdělení $F(v_1; v_2)$	v_1; v_2 – počet stupňů volnosti	= *FDIST*(x;v_1;v_2)	Rozdělení: F (Fischerovo), kde sv 1 je parametr v_1 , sv 2 je parametr v_2 a F je libovolné nezáporné číslo.

Tabulka 2. Přehled statistických funkcí diskrétních rozdělení.

2.2.4 Normální rozdělení (Gaussovo-Laplaceovo) N($\mu;\sigma^2$)

Normální rozdělení je ve statistické praxi nejčastěji zmiňovaným spojitým rozdělením. Řada statistických metod je založena na předpokladu, že analyzovaná proměnná nebo vypočtená charakteristika mají normální rozdělení. Při statistickém zpracování měření veličin v přírodních vědách, v technice i jinde, mívají chyby měření právě normální rozdělení. Řada jiných rozdělení přechází za určitých podmínek v rozdělení normální, takže jiná než normální rozdělení často nahrazujeme přibližně normálním rozdělením. I rozdělení některých výběrových charakteristik v náhodných výběrech modelujeme normálním rozdělením. Rozdělení je určeno dvěma parametry: střední hodnotou rozdělení μ a rozptylem σ^2. Normální rozdělení často značíme $N(\mu;\sigma^2)$. Pro střední hodnotu rozdělení tedy platí $E(X) = \mu$ a

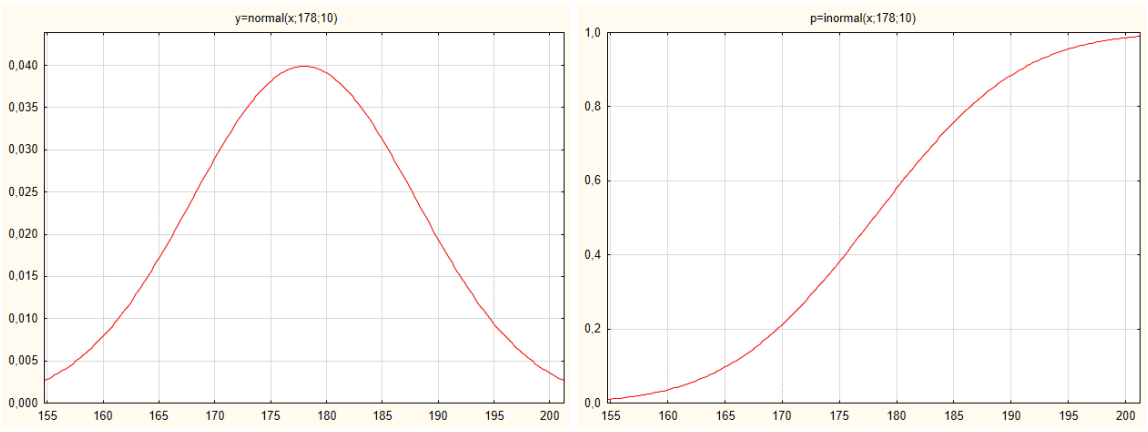
pro rozptyl $D(X) = \sigma^2$. Náhodná veličina X může nabývat všech hodnot od mínus do plus nekonečna.

Funkce *hustoty pravděpodobnosti* $f(x)$ je v literatuře používána k popisu spojité veličiny častěji než funkce distribuční. Pravděpodobnostní funkce spojité náhodné veličiny totiž není definovaná. U spojitých rozdělení lze pouze určit pravděpodobnost, že se objeví hodnota v určitém intervalu nikoliv hodnota samotná (tzv. paradox nulové pravděpodobnosti). Pravděpodobnost závisí na šíři sledovaného intervalu tak, že s jeho rozšířením roste nebo může zůstávat i stejná, ale nikdy nemůže klesat. Za účelem odstranění závislosti pravděpodobnosti na šíři intervalu podělíme pravděpodobnost výskytu hodnot z daného intervalu jeho šíří, takže vztáhneme tuto pravděpodobnost na jednotkovou šíři. Novou veličinu nazýváme hustota pravděpodobnosti. Grafické znázornění této funkce popisuje vlastnosti rozdělení lépe než distribuční funkce. Tu lze z hustoty pravděpodobnosti vypočítat integrálním počtem jako:

$$F(x_1) = \int_{-\infty}^{x_1} f(x)dx\,.$$

Funkce hustoty pravděpodobnosti normálního rozdělení má poněkud složitý vzorec:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,.$$



Obr. Hustota pravděpodobnosti a distribuční funkce pro normální rozdělení s parametry $\mu = 178$ a $\sigma = 10$.

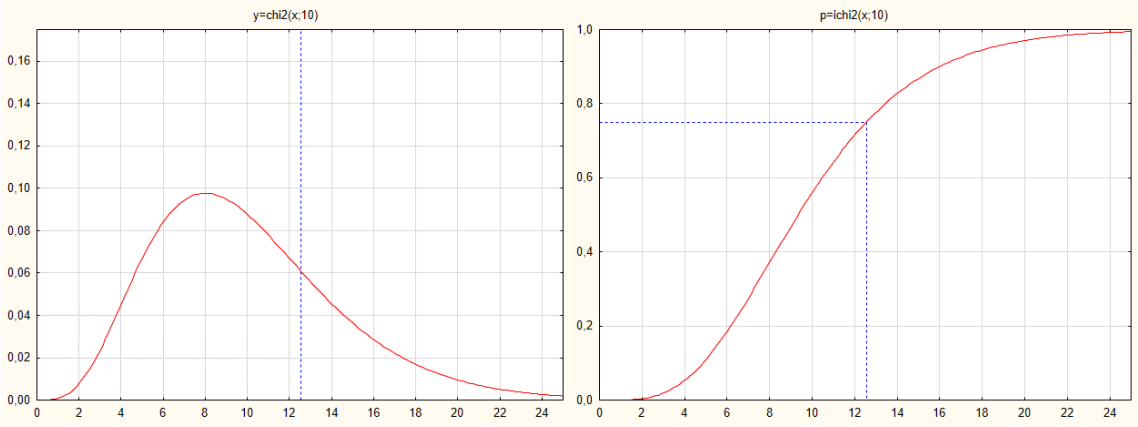
Specifickým případem normálního rozdělení je normované normální rozdělení s nulovou střední hodnotou a jednotkovým rozptylem, tedy N(0;1). Náhodná veličina se značí v tomto případě obvykle U (nebo Z).

Centrální limitní věta zdůvodňuje výsadní postavení normálního rozdělení. Existují její různé varianty. V principu jde o to, že pravděpodobnostní rozdělení náhodné veličiny, která vznikla jako součet nebo průměr velkého počtu vzájemně nezávislých nebo jen slabě závislých náhodných veličin, je možné za poměrně širokých podmínek nahradit (aproxi-

movat) normálním rozdělením, i když dílčí náhodné veličiny mají rozdělení různá. Rozdělení takové náhodné veličiny se totiž blíží k normálnímu rozdělení jako limitnímu, pokud je počet dílčích veličin dostatečně velký.

2.2.5 Rozdělení chí-kvadrát $\chi^2(v)$

Toto spojité rozdělení je odvozeno z normovaného normálního rozdělení U podle funkce $\chi^2 = (U_1^2 + U_2^2 + \ldots + U_v^2)$, kde U_i jsou náhodné veličiny s normovaným normálním rozdělením. Je to nesymetrické rozdělení zešikmené k vyšším hodnotám. Má jediný parametr v, který nazýváme *počet stupňů volnosti*. Střední hodnota rozdělení $E(X) = v$. Rozptyl rozdělení $D(X) = 2v$. Kvantily tohoto rozdělení se značí $\chi_p^2(v)$ a využívají se zejména při testování statistických hypotéz.



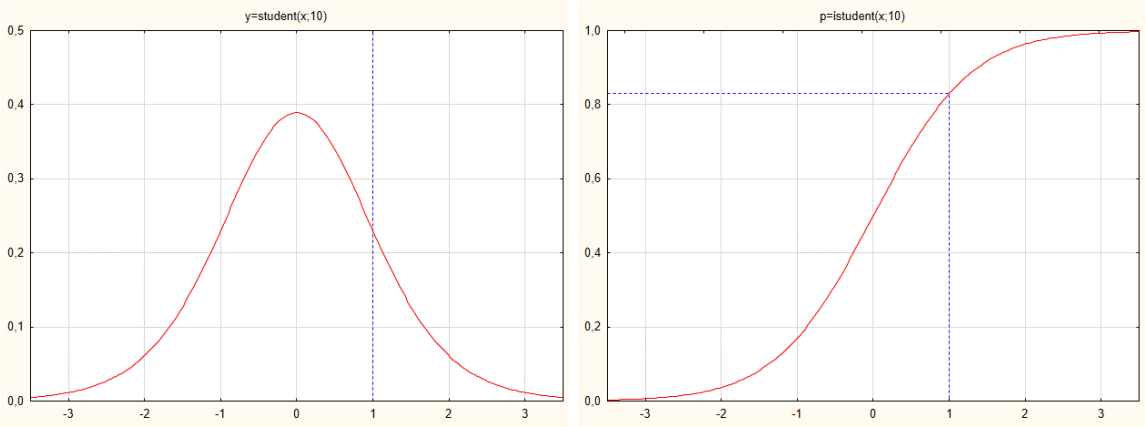
Obr. Hustota pravděpodobnosti a distribuční funkce chí-kvadrát rozdělení se stupni volnosti $v = 10$.

2.2.6 Studentovo rozdělení (t-rozdělení) t(v)

Toto rozdělení je odvozeno z veličin s normovaným normálním U a chí-kvadrát rozdělením podle funkce:

$$t = \frac{U}{\sqrt{\frac{\chi^2}{v}}},$$

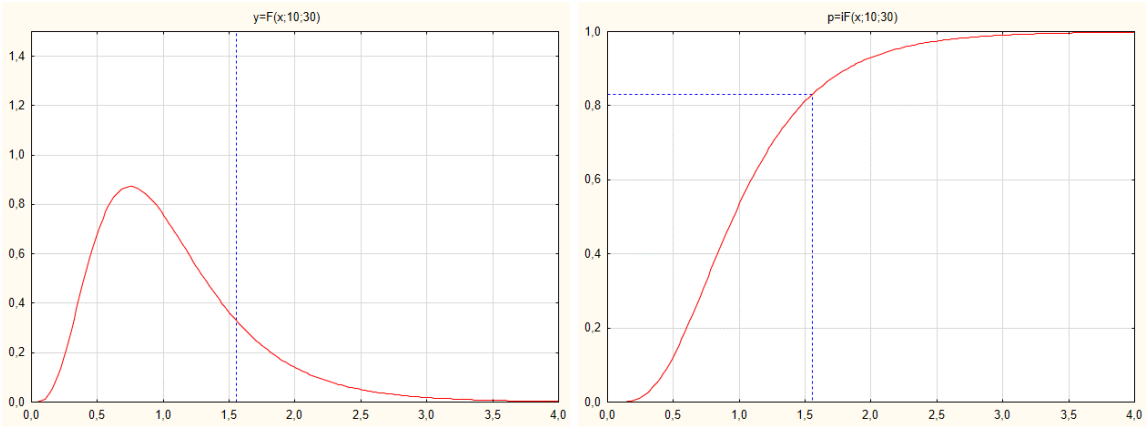
kde v jsou stupně volnosti chí-kvadrát rozdělení. Studentovo rozdělení je rozložené symetricky kolem nuly. Má jediný parametr v, který je opět nazýván počet stupňů volnosti. Podobá se tvarem normovanému normálnímu rozdělení, má však menší špičatost. Pro velké hodnoty parametru v přechází toto rozdělení v rozdělení normální. Co se charakteristik rozdělení týká, jsou střední hodnota, modus a medián rovny nule pro $v > 1$, jinak nejsou definovány. Rozptyl $D(X) = v/(v\text{-}2)$ pro $v > 2$, jinak není definován. Kvantily se značí $t_p(v)$ a využívají se zejména při testování statistických hypotéz.



Obr. Hustota pravděpodobnosti a distribuční funkce Studentova rozdělení se stupni volnosti $\nu = 10$.

2.2.7 F-rozdělení (Fisher-Snedecorovo) F(ν_1;ν_2)

Toto rozdělení vychází z funkce $F = (W/\nu_1)/(V/\nu_2)$, kde W a V jsou dvě nezávislé náhodné veličiny s rozděleními χ^2 s ν_1 a ν_2 stupni volnosti. Jedná se o asymetrické rozdělení kladně zešikmené. Přechází na normované normální rozdělení pro $\nu_1 = 1$ a $\nu_2 \to \infty$. Kvantily $F_p(\nu_1;\nu_2)$ se využívají zejména při testování statistických hypotéz. Střední hodnota $E(X) = \nu_2/(\nu_2\text{-}2)$ pro $\nu_2 > 2$ a rozptyl: $D(X) = \dfrac{2\nu_2^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-2)^2(\nu_2-4)}$ pro $\nu_2 > 4$.



Obr. Hustota pravděpodobnosti a distribuční funkce F-rozdělení se stupni volnosti $\nu_1 = 10$ a $\nu_2 = 30$.

2.3 Metody odhadu parametrů rozdělení základního souboru

Zobecňováním vlastností zjištěných pomocí výběru na celý základní soubor se zabývá *statistické usuzování*. Tento obor statistiky dělíme obvykle na:

1. *Teorie odhadů.* Předpokládáme určitý typ rozdělení základního souboru a z výsledků měření prvků výběru odhadujeme parametry (např. střední hodnotu, směrodatnou odchylku, pravděpodobnost, modus, medián) rozdělení základního souboru.

2. *Testování hypotéz.* Ověřování tvrzení o základním souboru, které vzniklo ještě před začátkem sběru dat.

Základní představa usuzování vychází z toho, že *základní soubor* je obvykle hypotetický, reálně neexistující a většinou i nekonečně velkého rozsahu a dále že jeho hodnoty se řídí určitým rozdělením pravděpodobnosti (rozdělením teoretickým). Daný základní soubor je vždy jen jeden. Skutečně naměřené hodnoty statistické proměnné jsou však získány na reálném a konečném výběru ze základního souboru. Takových výběrů z jednoho základního souboru může být více. *Výběrový soubor* je tedy proměnlivý.

Statistické usuzování vyžaduje kvalitní data. Výběr by měl být typickým vzorkem základního souboru, věrným a zmenšeným obrazem. Má být reprezentativní. Z toho důvodu by měl splňovat několik náležitostí:

1. Všechny prvky základního souboru mají stejnou pravděpodobnost, že budou zařazeny do výběru.
2. Výběr má být homogenní, tj. všechny prvky mají pocházet ze stejného souboru, tedy ze stejného rozdělení pravděpodobnosti.
3. Jednotlivé prvky výběru mají být nezávislé. Nesmí vznikat chyba vzájemným ovlivňováním vzorků nebo chyba závislá na pořadí měření nebo na čase. Nesmí se ani měnit podmínky během odběru.

2.3.1 Bodové odhady

Parametr je číselná hodnota popisující určitou vlastnost základního souboru, považujeme jej za konstantu. Protože tento nedokážeme změřit z různých důvodů, zůstává pro nás tato hodnota neznámá. Můžeme však spočítat jeho odhad z výběru z populace pomocí výběrové charakteristiky. K rozlišení se používá rozdílné značení. Parametry se většinou značí písmeny řecké abecedy (průměr μ, rozptyl σ^2, medián $\tilde{\mu}$, modus $\hat{\mu}$, pravděpodobnost π atd.). Výběrové charakteristiky se značí písmeny latinské abecedy (průměr $\bar{x}$, rozptyl s^2, medián $\tilde{x}$, modus $\hat{x}$, pravděpodobnost p). Pokud odhadujeme parametr základního souboru jedním číslem, jde o bodový odhad. Aby byl odhad vhodný a použitelný pro praxi, měl by být:

1. *Konzistentní* - s rostoucím rozsahem výběru n se odhad blíží k parametru s pravděpodobností 1.
2. *Nestranný* (nevychýlený, nezkreslený) - odhad nesmí soustavně nadhodnocovat či podhodnocovat odhadovaný parametr.
3. *Vydatný* - rozptylu odhadů okolo odhadovaného parametru při opakovaných výběrech je malý. Ze dvou různě počítaných odhadů (různé vzorečky) je vydatnější ten, který má menší rozptyl.
4. *Rezistentní* - není příliš závislý na odchylkách od předpokládaného rozdělení.

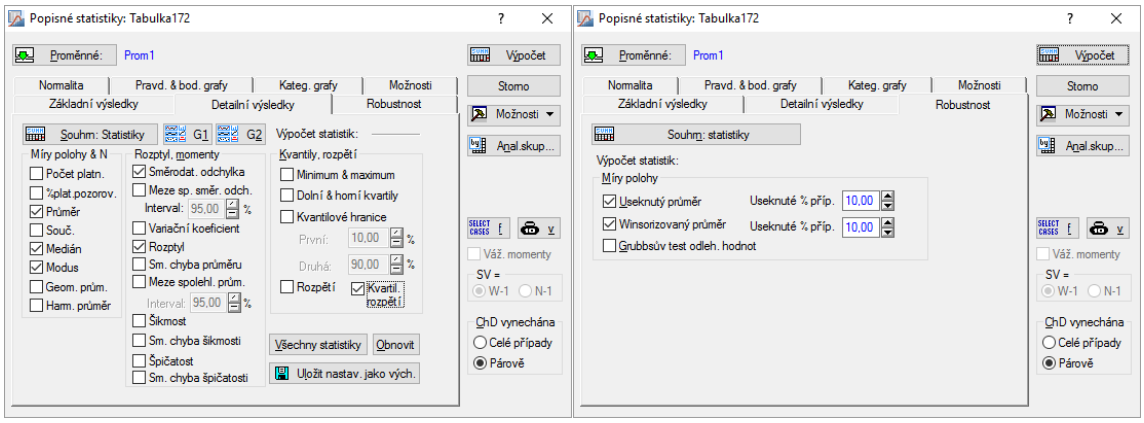
V případě odhadu průměru μ je aritmetický průměr jeho nejvydatnějším odhadem, méně vydatný je výběrový medián a nejmenší hodnotu vydatnosti má průměr počítaný z extrémních hodnot souboru. Rezistentním odhadem průměru μ je výběrový medián. Nestranným odhadem průměru μ jsou aritmetický průměr a medián pouze tehdy, pokud má rozdělení náhodné veličiny symetrický tvar. Nejvydatnějším odhadem rozptylu σ^2 je výběrový rozptyl s^2.

V případě, kdy se ve výběrovém souboru nacházejí odlehlé nebo extrémní hodnoty nemusejí výše uvedené charakteristiky správně plnit svou funkci. *Robustní charakteristiky* jsou takové, u kterých vychýlené hodnoty nemají vliv na kvalitu odhadu. Robustní charakteristiky typicky používáme v případech, když:

1. Výběrový soubor má odlehlé nebo z odlehlosti podezřelé výsledky, které nemohly být opraveny a není vhodné je vyloučit.
2. Výběrový soubor nemá normální rozdělení výsledků, např. vykazuje významnou šikmost.
3. Hodnocený soubor má velké rozptýlení dat.

Robustním odhadem střední hodnoty může být medián, useknutý průměr nebo Winsorizovaný průměr. Useknutý průměr se počítá jako průměr aritmetický s tím, že je před výpočtem odstraněno určité stejné procento nejvyšších a nejnižších hodnot. Obdobný princip využívá i Winsorizovaný průměr, který spočívá v náhradě stejného procenta nejvyšších a nejnižších hodnot v uspořádaném souboru dat hodnotou sousední. Tzn. první hodnota se nahradí druhou, *n*-tá hodnoty (*n*-1)-ní. Tímto postupem se do určité míry omezí odlehlost dat, ale zachová se jejich trend – velký výsledek se nahradí opět velkým a malý malým. Robustním odhadem směrodatné odchylky může být například charakteristika $s_R = 0,7413\left(\tilde{x}_{0,75} - \tilde{x}_{0,25}\right)$, kde $\tilde{x}_{0,75}$ - $\tilde{x}_{0,25}$ je kvartilové rozpětí (rozdíl mezi horním a dolním kvartilem souboru). Z ní lze pak odvodit odhad rozptylu jako s_R^2.

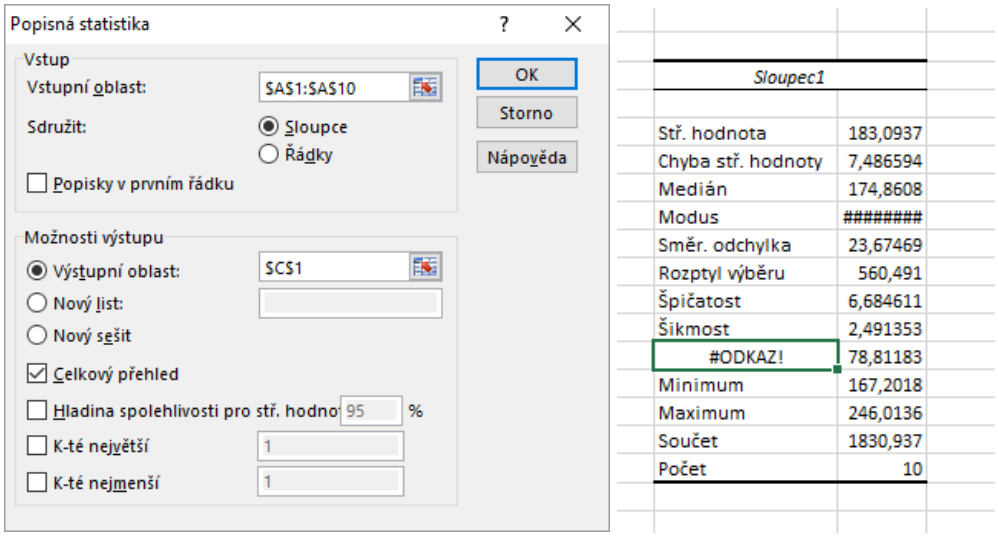
Základní charakteristiky souboru lze v programu Statistika spočítat v nabídce Statistiky → Základní statistiky/tabulky → Popisné statistiky. Na záložce Detailní výsledky lze zaškrtnout požadované charakteristiky. Useknutý a Winsorizovaný průměr program počítá také ve stejné nabídce na záložce Robustnost. U obou je nutno zadat požadované procento useknutých resp. nahrazených hodnot souboru. Pokud např. vložím 10 %, bude ze souboru odstraněno resp. nahrazeno 10 % nejnižších a 10 % nejvyšších hodnot. Robustní charakteristiku s_R program nepočítá, ale na kartě Detailní výsledky lze nastavit výpočet kvartilového rozpětí a z něj lze s_R jednoduše dopočítat.



Obr. Zadávací okno nástroje Popisné statistiky programu Statistika.

V programu MS Excel lze charakteristiky spočítat pomocí funkcí. Dobrý přehled (průměr, medián, modus, rozptyl, směrodatná odchylka, kvartilové rozpětí a další) nabízí i nástroj Popisná statistika (Data → Analýza dat → Popisná statistika). Odkaz na data se vkládá do políčka Vstupní oblast a je nutno zaškrtnout volbu Celkový přehled. Useknutý průměr je nutno počítat funkcí TRIMMEAN. První argument je oblast dat a druhý procento useknutých hodnot. Naproti programu Statistica se zde zadává celkové procento useknutých hodnot. Pokud např. vložím 20 % bude ze souboru odstraněno 10 % nejnižších a 10 % nejvyšších

hodnot, tj. funkce má tvar =TRMMEAN(A1:A10;20%). Jiné funkce pro robustní charakteristiky program nemá. Charakteristiku s_R lze spočítat s využitím funkcí percentil jako =0,7413*(PERCENTIL(A1:A10;0,75)- PERCENTIL(A1:A10;0,25)).



Obr. Zadávací okno nástroje Popisná statistika programu MS Excel a výstup.

Příklad: Pokračujeme s daty o výšce lidí, kdy ze základního souboru výšek lidí (předpokládáme normální rozdělení se střední hodnotou 178 cm, rozptylem 100 a směrodatnou odchylkou 10 cm) provedeme náhodný výběr 10 lidí. Data jsou následující:
167,20; 167,91; 170,50; 171,06; 172,76; 176,96; 179,07; 184,73; 194,74; 246,01
V souboru je jedna hodnota podezřelá jako odlehlá (246,01 cm), která by mohla ovlivnit kvalitu odhadů parametrů základního souboru. Z vypočtených charakteristik je patrné, že ty robustní se parametrům základního souboru blíží více, než charakteristiky základní. Vypočtené charakteristiky jsou následující:

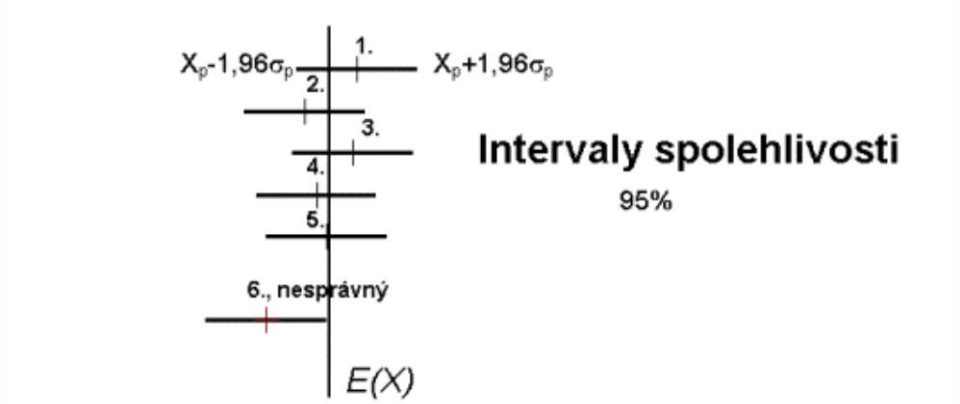
Základní charakteristiky	Robustní charakteristiky
Průměr = 183,09 cm	Medián = 174,86 cm Useknutý průměr(10%) = 177,22 cm Winsorizovaný průměr(10%) = 178,04 cm
Rozptyl = 560,49 Směrodatná odchylka = 23,68 cm	$s_R^2 = (0{,}7413*14{,}23)^2 = 111{,}28$ $s_R = 0{,}7413*14{,}23 = 10{,}55$ cm

2.3.2 Intervalové odhady

Bodový odhad nás zdánlivě naplňuje jistotou přesně stanoveného čísla, které nám umožňuje bez problémů s tímto odhadem pracovat. Například jej lze srovnávat s nějakým předepsaným limitem. Opak je však pravdou, protože bodový odhad se prakticky nikdy nemůže „trefit“ do odhadované hodnoty a při opakovaném určení odhadu z jiného výběru dostaneme odlišnou hodnotu bodového odhadu. Není-li udán interval vyjadřující možné rozptýlení hodnot odhadu, může být jeho zamlčená nejistota tak veliká, že odhad je pro náš účel prakticky bezcenný.

Ze statistického pohledu se jako vhodnější jeví odhad intervalový, kdy sestrojujeme určité rozmezí parametru - *interval spolehlivosti*. Jeho šířka závisí na *hladině spolehlivosti*. Hladina spolehlivosti je pravděpodobnost, se kterou interval spolehlivosti pokryje parametr základního souboru při opakovaném provádění výběru. Nejpoužívanější hladiny spolehlivosti jsou 90 %, 95 % nebo 99 %. Pokud např. použiji hladinu spolehlivosti 95 %, zname-

ná to, že ze 100 vypočtených intervalů spolehlivosti jich přibližně 95 pokryje hodnotu parametru. Interval spolehlivosti tak obdobně jako bodový odhad má náhodně proměnlivý charakter. Každý výběr vede k trochu jinému intervalu spolehlivosti, jak je uvedeno na následujícím obrázku pro 6 náhodných výběrů ze stejného základního souboru.



Obr. Porovnání intervalů spolehlivosti pro odlišné výběry ze stejného základního souboru.

2.3.3 Intervalový odhad střední hodnoty a rozptylu

Za podmínky, že výběrový soubor je malý, ale základní soubor je normálně rozdělen nebo pokud je výběrový soubor velký (v praxi $n > 30$), lze spočítat intervaly spolehlivosti s využitím kvantilů normovaného normálního rozdělení. Připomeňme, že podle centrální limitní věty mají výběrové průměry přibližně normální rozdělení i v případě jiného než normálního rozdělení základního souboru a to pokud je počet prvků výběru n dostatečně velký.

Při výpočtu intervalu spolehlivosti většinou využijeme výpočet založený na předpokladu, že neznáme směrodatnou odchylku rozdělení σ. Tu, stejně jako odhadujeme střední hodnotu výběrovým průměrem $\bar{x}$, musíme odhadnout výběrovou směrodatnou odchylkou s. Charakteristika:

$$t = \frac{\bar{x} - \mu}{s}\sqrt{n}$$

má Studentovo rozdělení s $v = n$-1 stupni volnosti, což je jediný parametr tohoto rozdělení. Oboustranný (1-α)*100% interval spolehlivosti je dán rovnicí:

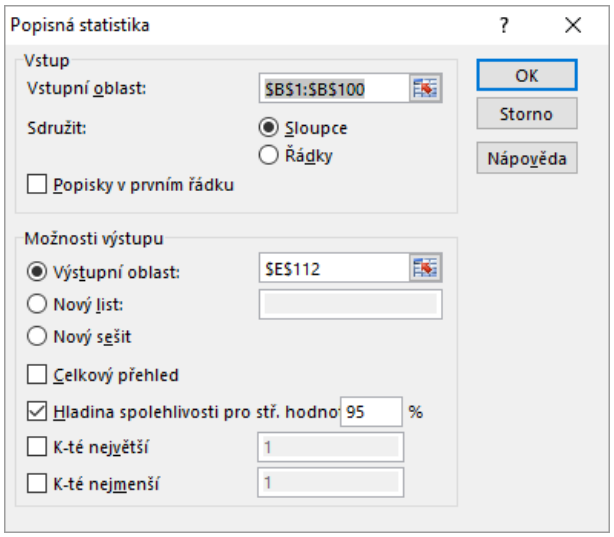
$$P\left(\bar{x} - t_{1-\alpha/2}(n-1)\cdot\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{1-\alpha/2}(n-1)\cdot\frac{s}{\sqrt{n}}\right) = 1-\alpha\,.$$

Interval spolehlivosti pro rozptyl je dán rovnicí:

$$P\left(\frac{(n-1)\cdot s^2}{\chi^2_{1-\alpha/2}(n-1)} \le \sigma^2 \le \frac{(n-1)\cdot s^2}{\chi^2_{\alpha/2}(n-1)}\right) = 1-\alpha\,,$$

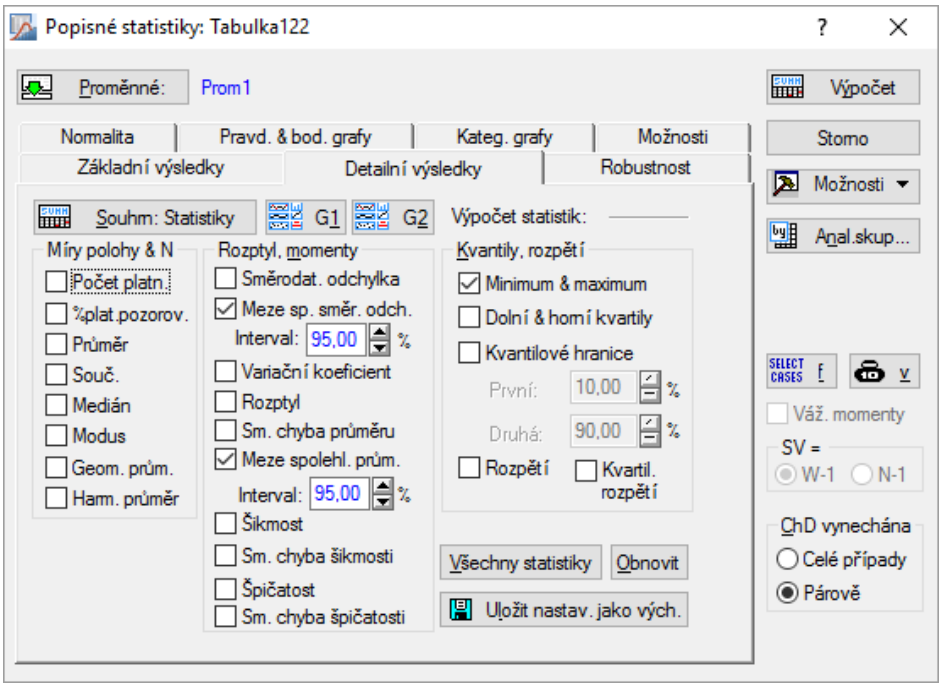
kde $\chi^2_{1-\alpha/2}(n\text{-}1)$ a $\chi^2_{\alpha/2}(n\text{-}1)$ jsou kvantily rozdělení χ^2 s n-1 stupni volnosti. Odmocněním mezí pak získáme interval spolehlivosti pro směrodatnou odchylku.

Interval spolehlivosti pro střední hodnotu lze spočítat jak v MS Excel, tak i v programu Statistica. V programu MS Excel můžeme využít funkci CONFIDENCE.T. Pozor, ta je implementována od verze 2013. Starší verze mají funkci CONFIDENCE, ale ta počítá interval za poněkud jiných podmínek a dle jiného vzorce. Pomocí zmíněné funkce spočítáme hodnotu $\delta = t_{1-\alpha/2}(n-1)\cdot\frac{s}{\sqrt{n}}$. Funkce vyžaduje tři argumenty v následujícím pořadí: α; výběrovou směrodatnou odchylku s; počet hodnot výběru n. Dolní mez intervalu spolehlivosti pak jednoduše dopočítáme jako $\bar{x}-\delta$ a horní mez jako $\bar{x}+\delta$. Alternativně lze bez ohledu na verzi programu použít nástroj Popisná statistika (Data → Analýza dat → Popisná statistika) a hodnotu δ spočítat zaškrtnutím položky Hladina spolehlivosti pro stř. hodnotu. Do vstupního políčka se zadává požadovaná hladina spolehlivosti v procentech. Tento nástroj vyžaduje zadat výběrová data do pole Vstupní oblast. Program nenabízí výpočet intervalů pro jiné parametry než je střední hodnota. Ty je nutno počítat dle vzorců.



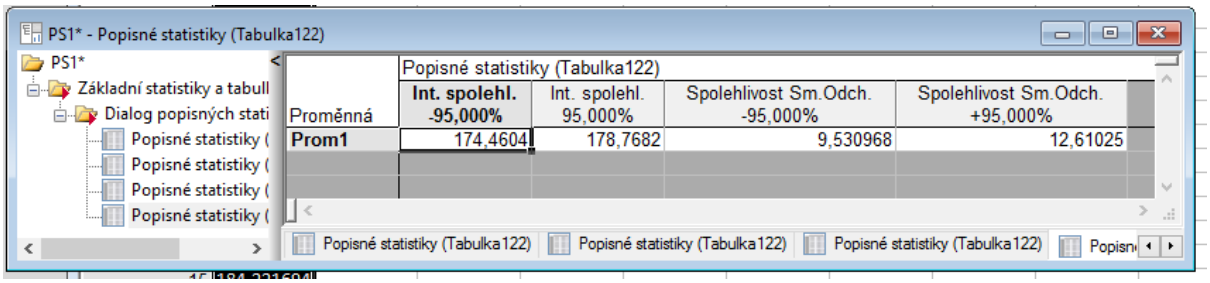
Obr. Zadávací okno nástroje Popisná statistika programu MS Excel.

Intervaly spolehlivosti pro střední hodnotu i směrodatnou odchylku lze spočítat v programu Statistica v nabídce Statistiky → Základní statistiky/tabulky → Popisné statistiky. Na záložce Detailní výsledky je nutno zaškrtnout volbu Meze spolehli. prům. (pro střední hodnotu) a Meze sp. směr. odch. (pro směrodatnou odchylku) a u obou nastavit požadovanou hladinu spolehlivosti v procentech.



Obr. Zadávací okno nástroje Popisné statistiky programu Statistika.

Příklad: Pokračujeme s datovým souborem výšky lidí, kdy ze základního souboru výšek lidí (předpokládáme normální rozdělení se střední hodnotou 178 cm a směrodatnou odchylkou 10 cm) provedeme náhodný výběr 100 lidí.
Bodový odhad střední hodnoty provedeme pomocí aritmetického průměru, který je jeho nestranným a vydatným odhadem a vyšel zaokrouhleně 176,61 cm. Bodový odhad směrodatné odchylky provedeme pomocí výběrové směrodatné odchylky, která vyšla zaokrouhleně 10,86 cm. V programu MS Excel jsme zadali funkci =CONFIDENCE.T(0,05;10,86;100) a vyšla hodnota δ = 2,16. Dolní mez 95% intervalu spolehlivosti má hodnotu 176,61-2,16=174,46 cm a horní mez 176,61+2,16=178,77 cm. Skutečný parametr má hodnotu 178 cm a je patrné, že interval spolehlivosti ji pokryl. Pokud bychom v programu použili nástroj Popisná statistika, vyšla by hodnota δ stejně.
Stejné hodnoty mezí získáme i programem Statistica s tím, že program přímo počítá meze jak pro střední hodnotu (Int. Spolehl.), tak i pro směrodatnou odchylku (Spolehlivost Sm. Odch.). Protože se jedná o stejný způsob výpočtu, vyšel interval spolehlivosti stejně jako v programu MS Excel.



Obr. Výstupy z nástroje Popisné statistiky programu Statistica.

2.3.4 Intervalové odhady metodou bootstrap

Existují i alternativní způsoby výpočtu intervalů spolehlivosti používané v situacích, kdy nejsou splněny výše uvedené podmínky velikosti výběru nebo rozdělení základního souboru. Jedním z nich je metoda *bootstrap*. Tato metoda je založena na myšlence odhadování parametru z velkého počtu náhodných výběrů. Protože takový postup často v praxi není možný, využívá metoda výběrového souboru a z něj vytváří náhodným výběrem s opakováním nové výběry stejného rozsahu. Takto získaný náhodný výběr se nazývá *bootstrapový výběr* nebo *bootstrapový soubor*. Bootstrapový výběr se mnohokrát opakuje (*B*-krát), přičemž platí, že $B >> n$. Počet všech možných bootstrapových výběrů je:

$$\binom{2n-1}{n}.$$

Metoda bootstrapových intervalových odhadů je užitečná, potřebujeme-li určit intervalové odhady parametrů pozorované náhodné veličiny, ale neznáme nebo nejsme schopni odhadnout rozdělení pravděpodobnosti dané veličiny a rozsah výběru není dostatečně velký, abychom mohli aplikovat asymptotické odhady. Bootstrapový odhad neslouží přímo k odhadu parametru samotného, ale používá se mj. právě při intervalových odhadech.
Obecný postup metody je následující:

1. Získání původního statistického souboru.
2. Výpočet charakteristik původního statistického souboru.
3. Vytvoření boostrapových výběrů.
4. Výpočet bootstrapových charakteristik.
5. Výpočet bootstrapových intervalových odhadů.

Existuje řada způsobů jak z bootstrapových výběrů vypočítat meze intervalů spolehlivosti. *Jednoduchý intervalový odhad* má tvar:

$$P\left(2t - t^*_{1-\alpha} \le t \le 2t - t^*_{\alpha}\right) = 1-\alpha,$$

kde t je libovolná výběrová charakteristika (např. průměr, směrodatná odchylka apod.) a $t^*_{1-\alpha}$ a t^*_{α} jsou kvantily vypočtených bootstrapových charakteristik. Početně jednoduchý je i *kvantilový interval*, který je založen na samotných kvantilech bootstrapových statistik a počítá se jako:

$$P\left(t^*_{\alpha/2} \le t \le t^*_{1-\alpha/2}\right) = 1-\alpha.$$

Velmi dobrých výsledků dosahují intervaly počítané metodou BCa. Jejich výpočet je však složitý a vyžaduje speciální software. Metoda bootstrap je početně náročná z důvodů generování velkého počtu bootstrapových výběrů. V Excelu ji však lze provést překvapivě jednoduše, jak ukáže následující příklad.

Příklad: Pokračujeme s daty o výšce lidí, kdy ze základního souboru výšek lidí (předpokládáme normální rozdělení se střední hodnotou 178 cm a směrodatnou odchylkou 10 cm) provedeme náhodný výběr 10 lidí, tedy získáme velmi malý soubor.
Bootstrapový odhad vytvoříme pomocí funkcí INDEX a RANDBETWEEN. Funkce INDEX vybírá z předem definované oblasti jednu hodnotu. Má tři argumenty: oblast, kde se nachází výběrový soubor; řádky, ze kterých má vybírat; sloupce, ze kterých má vybírat. Vzhledem k tomu, že proměnná je v jednom sloupci je pro bootstrapový výběr potřeba náhodně vybrat z řádků, a proto bude v druhém argumentu použita funkce RANDBETWEEN se dvěma argumenty: první řádek výběru; poslední řádek výběru. Výběrová data jsou umístěna ve sloupci A na řádcích 2 až 11, tedy funkce má tvar =INDEX($A:$A;RANDBETWEEN(2;11);1). Takto zapsaná funkce náhodně vybere jednu hod-

notu z výběru. Bootstrapový výběr má stejný rozsah jako původní výběr, tedy tuto funkci zkopírujeme přes deset sousedních políček v jednom řádku. Tím jsme získali první bootstrapový výběr. Celou funkci zkopírujeme přes 999 dalších řádků a získáme 1000 bootstrapových výběrů. Celkem lze ze souboru 10 čísel získat 92378 rozdílných bootstrapových výběrů.
Pro každý bootstrapový výběr spočítáme průměr do sloupce M a získáme tak soubor 1000 průměrů. Pro výpočet mezí jednoduchou metodou pak dolní mez počítáme jako =2*PRŮMĚR(A2:A11)-PERCENTIL(M2:M1001;0,95) a horní mez pak jako =2*PRŮMĚR(A2:A11)-PERCENTIL(M2:L1001;0,05). Výsledky jsou uvedeny v přehledné tabulce na konci příkladu. Pokud bychom chtěli počítat meze kvantilovou metodou aplikujeme funkce percentil =PERCENTIL(M2:M1001;0,025) pro výpočet dolní meze a =PERCENTIL(M2:M1001;0,975) pro výpočet horní meze 95% intervalu spolehlivosti.

Při počítání bootstrapového intervalu spolehlivosti pro směrodatnou odchylku využijeme 1000 vytvořených bootstrapových výběrů a pro každý vypočítáme výběrovou směrodatnou odchylku funkcí SMODCH.VÝBĚR. Intervaly pak vypočteme analogicky, přičemž u jednoduchého typu nahradíme ve výpočtu výběrový průměr právě výběrovou směrodatnou odchylkou. Takovýmto způsobem můžeme odhadovat i další parametry základního souboru (např. medián a šikmost). Bootstrapové intervaly jsou užší než asymptotické a nejsou symetrické kolem bodového odhadu.

		Bodový odhad	**Asymptotický interval v MS Excel**	**Asymptotický interval ve Statistica**	**Bootstrapový interval (jednoduchý)**	**Bootstrapový interval (kvantilový)**
Střední hodnota základního souboru (178 cm)	Dolní mez	174,81	168,09	168,09	169,77	169,87
	Horní mez		181,54	181,54	179,17	180,57
Směrodatná odchylka základního souboru (10 cm)	Dolní mez	9,40	-	6,47	6,74	4,35
	Horní mez		-	17,16	14,08	12,58
Medián základního souboru (178 cm)	Dolní mez	171,91	-	-	165,08	167,91
	Horní mez		-	-	174,62	179,07
Šikmost základního souboru (1)	Dolní mez	1,09	-	-	0,23	-0,42
	Horní mez		-	-	2,41	2,20

Tab. Porovnání asymptotických a bootstrapových intervalů spolehlivosti

Příklad: Základní soubor dat má zešikmené chí-kvadrát rozdělení se stupni volnosti v=8. Z tohoto souboru jsme provedli náhodný výběr 20 hodnot. Porovnáme asymptotické a bootstrapové intervaly spolehlivosti. Opět platí, že bootstrapové intervaly jsou užší než asymptotické a nejsou symetrické kolem bodového odhadu

		Bodový odhad	Asymptotický interval v MS Excel	Asymptotický interval ve Statistica	Bootstrapový interval (jednoduchý)	Bootstrapový interval (kvantilový)
Střední hodnota základního souboru (8)	Dolní mez	7,85	6,86	6,89	6,39	6,13
	Horní mez		8,95	9,82	9,31	9,68
Směrodatná odchylka základního souboru (4)	Dolní mez	4,20	-	3,12	3,25	2,58
	Horní mez		-	6,14	5,57	5,33
Medián základního souboru (7,34)	Dolní mez	6,62	-	-	3,34	4,44
	Horní mez		-	-	8,77	9,94

Tab. Porovnání asymptotických a bootstrapových intervalů spolehlivosti

2.4 Testování hypotéz

Významnou a ve výzkumu často používanou formou statistického usuzování je *testování hypotéz*. Jeho cílem je najít odpověď ve formě ano/ne na předem položenou otázku. Proceduru testování lze rozložit do následujících kroků:

1. Formulace výzkumné otázky ve formě *nulové* a *alternativní hypotézy*. Ty zjednodušeně představují odpovědi ano a ne, přičemž jejich přiřazení k hypotézám závisí na typu testu. Nulová hypotéza (značíme H_0) obvykle vyjadřuje tvrzení „žádný rozdíl“ (tj. jakýkoliv rozdíl nalezený v datech lze přičíst jejich variabilitě). Alternativní hypotéza (značíme H_A nebo H_1) pak znamená situaci, kdy H_0 neplatí. To pak odpovídá existenci diference nebo závislosti dle typu výzkumné otázky. Každý test se chová jako soudce. Platí princip presumpce neviny, tedy přijímá H_0 a pouze pokud je dostatek důkazů je hypotéza H_0 zamítnuta. Nezamítnutí nulové hypotézy ale neznamená její důkaz. Spíše jde o zjištění, že jsme nenalezli dostatek prostředků k jejímu zamítnutí.
2. Volba přijatelné chyby rozhodování, tedy volba tzv. *hladiny významnosti* α, což je pravděpodobnost, že test povede k zamítnutí nulové hypotézy, ačkoliv tato platí. Hodnota hladiny významnosti se v praxi volí velmi malá 0,1; 0,05 nebo 0,001.
3. Výpočet *testovací statistiky*, která je základem pro provedení úvah o platnosti hypotéz. V závislosti na povaze dat a formulovaných hypotézách volíme testovací statistiku. Její volba je stěžejní pro správný výsledek testu. Pro testování střední hodnoty a v mnoha dalších případech se používá testovací statistika stanovaná jako standardizovaná odchylka odhadu od hypotetické hodnoty stanovené nulovou hypotézou a její obecný tvar je:

$$\text{testovací statistika} = \frac{\text{bodový odhad - hypotetická hodnota}}{\text{směrodatná chyba odhadu}}.$$

4. Při formulaci závěru testu rozhodneme o platnosti nebo zamítnutí nulové hypotézy. To lze provést dvěma způsoby. Buď porovnáme testovací statistiku s kritickou mezí. Ta je stanovena jako hranice, za kterou je hodnota testovací statistiky příliš ex-

trémní na to, aby mohla platit H_0. Mez je stanovena v závislosti na hladině významnosti α a její hodnoty jsou definovány pravděpodobnostními rozděleními nebo tabulkovými hodnotami v závislosti na typu testové statistiky. Statistické programy nabízejí často tzv. *p-hodnotu* testu, která kvantifikuje pravděpodobnost výskytu hodnoty testovací statistiky za podmínky, že nulová hypotéza platí. Jestliže je malá, pak je to doklad, že nulová hypotéza neplatí. P-hodnota se tedy porovnává s hladinou významnosti α. Pokud je menší, pak zamítáme H_0, jinak H_0 nezamítáme.

2.4.1 Testy o parametrech základního souboru

V následujících kapitolách se budete setkávat s řadou různých testů. Mezi nejčastěji používané patří testy o parametrech rozdělení základního souboru. Konkrétně se nyní budeme zabývat jeho střední hodnotou. Za podmínky, že výběrový soubor je malý, ale základní soubor je normálně rozdělen nebo pokud je výběrový soubor velký (v praxi $n > 30$) lze *využít t-test na střední hodnotu*. Jeho nulová hypotéza říká, že střední hodnota základního souboru μ má určitou konkrétní hypotetickou hodnotu μ_0. Alternativní hypotéza je jejím opakem, tedy lze zapsat:

$$H_0: \mu = \mu_0 \text{ vs. } H_A: \mu \neq \mu_0.$$

Test, který ověří platnost takto zapsané nulové hypotézy, se nazývá *oboustranný*. Vedle něj lze stanovit i jiné hypotézy. *Pravostranný* test bude ověřovat hypotézu:

$$H_0: \mu \leq \mu_0 \text{ vs. } H_A: \mu > \mu_0,$$

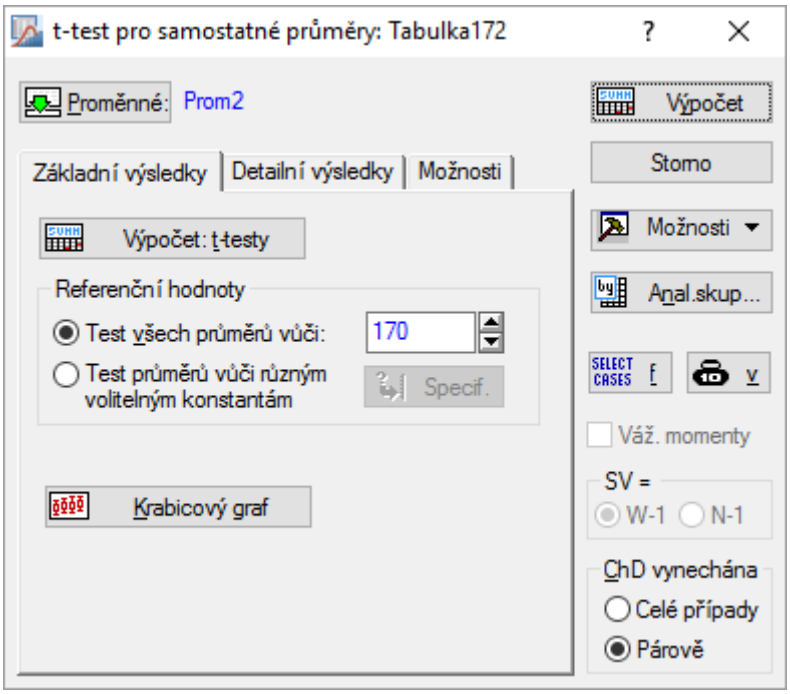
levostranný pak:

$$H_0: \mu \geq \mu_0 \text{ vs. } H_A: \mu < \mu_0.$$

Testovací statistika má tvar známý již z intervalů spolehlivosti:

$$t = \frac{\bar{x} - \mu}{s}\sqrt{n}$$

a stejně tak má Studentovo rozdělení s $\nu = n - 1$ stupni volnosti, což je informace, kterou využijeme při konstrukci kritického oboru testu nebo při výpočtu jeho p-hodnoty. V programu MS Excel lze p-hodnotu testu vypočítat funkcí ZTEST. Argumenty funkce jsou oblast dat; hypotetická střední hodnota. Vypočtená p-hodnota je platná pro pravostranný test. Pokud bychom prováděli levostranný test, je nutno p-hodnotu dopočítat jako *1 - p-hodnota pravostranného testu.* Pro oboustranný test platí vzorec *2*(menší z hodnot p-hodnota pravostranného testu a p-hodnota levostranného testu)*. Oboustrannou variantu testu počítá Statistica v nabídce Statistiky → Základní statistiky/tabulky → t-test, samost. vzorek. Do políčka Test všech průměrů vůči: se zadává hypotetická střední hodnota.



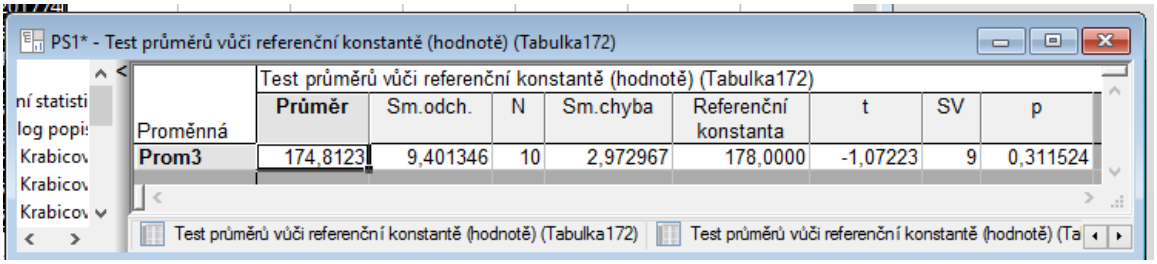
Obr. Zadávací okno nástroje t-test, samost. vzorek programu Statistika.

Příklad: Pokračujeme s daty o výšce lidí, kdy ze základního souboru výšek lidí (předpokládáme normální rozdělení se střední hodnotou 178 cm a směrodatnou odchylkou 10 cm) provedeme náhodný výběr 10 lidí. Ověříme hypotézu, že střední hodnota souboru je 178 cm. Pro připomenutí je bodový odhad získaný aritmetickým průměrem 174,81 cm. Testem zjistíme, zda je zjištěný rozdíl významný, nebo zda je způsoben pouze variabilitou v datech výběrového souboru. Při výpočtu budeme následovat výše uvedené kroky:

1. Formulace výzkumné otázky ve formě *nulové* a *alternativní hypotézy*.

$$H_0: \mu = 170 \text{ cm vs. } H_A: \mu \neq 170 \text{ cm},$$

   tedy půjde o oboustranný test.
2. Volba hladiny významnosti α. Standardní hodnota při testování je $\alpha = 0{,}05$.
3. Výpočet testovací statistiky t provedeme pomocí počítačových programů společně s p-hodnotou testu. V programu MS Excel jsou data v buňkách A1 až A10, tedy funkce má tvar `=ZTEST(A1:A10;178)`. Vypočtená p-hodnota 0,844 platí pro pravostranný test, abychom získali p-hodnotu pro test oboustranný, provedeme přepočet 2*(1-0,844) = 0,312. Stejný výsledek poskytne i program Statistica, který přímo počítá p-hodnotu oboustranného testu a spočítal i testovou statistiku t = -1,07233.



Obr. Výstupy z nástroje t-test, samost. vzorek programu Statistica.

4. K formulaci závěru testu využijeme p-hodnotu testu a ve druhém kroku stanovenou hladinu významnosti α. Protože p-hodnota je větší než ve druhém kroku zvolená hladina významnosti, nezamítáme H_0. Na hladině významnosti 0,05 tak platí, že střední hodnota základního souboru je 178 cm. Rozdíl mezi bodovým odhadem a skutečnou střední hodnotou tak byl dán pouze variabilitou v datech.

Pro názornost ještě provedeme testy pro další hypotetické střední hodnoty.

Hypotetická střední hodnota	p-hodnota testu	Závěr testu
150 cm	0,000	Zamítáme Ho
178 cm	0,311	Nezamítáme Ho
180 cm	0,115	Nezamítáme Ho
190 cm	0,001	Zamítáme Ho

Může být překvapivé, že i hodnota 180 cm je dle testu platná. To je však v souladu s principy statistického usuzování. Z výběrového souboru nelze jednoznačně určit jen jednu jedinou hodnotu parametru. Můžeme určit skupinu pravděpodobných hodnot parametru definovanou konkrétním rozmezím. Toto rozmezí se nazývá interval spolehlivosti a detailně jsme se mu věnovali v jedné z předchozích kapitol. Testy hypotéz a intervaly spolehlivosti jsou spolu úzce spjaty. Platí, že nulová hypotéza nebude zamítnuta pro žádnou hodnotu nacházející se uvnitř intervalu spolehlivosti a to na doplňkové (1- α) hladině významnosti jako je hladina spolehlivosti intervalu. Pro připomenutí vyšly meze asymptotického intervalu spolehlivosti pro střední hodnotu <168,09; 181,54>.

2.4.2 Neparametrické testy

Klasické postupy statistického usuzování obvykle předpokládají normální rozdělení základního souboru nebo využívají centrální limitní věty pro dostatečně rozsáhlé soubory. V praxi však tyto předpoklady často nejsou splněny, a proto je třeba podívat se po jiných typech testů, které jsou nezávislé na rozdělení základního souboru. *Neparametrické testy* jsou vhodné pro malé soubory, pro hodnocení ordinálních dat, nebo dat naměřených v poměrovém nebo intervalovém měřítku, jež nemají normální rozdělení. Využívají se také v případech, kdy je výběrový soubor zatížen odlehlými hodnotami. Existují různé typy neparametrických testů:

1. První typ pracuje namísto hodnot se znaménky odchylek od určité hypotetické hodnoty. Tímto postupem testuje střední hodnotu *znaménkový test hodnoty mediánu*.
2. Druhý typ nahrazuje původní hodnoty jejich pořadími, která jsou přiřazena po setřídění původních hodnot podle velikosti. *Wilcoxonův test hodnoty mediánu* je takovým testem střední hodnoty.
3. Třetí typ pracuje s původními daty a používá i stejné testovací statistiky jako v parametrických testech, ale jinak počítá p-hodnoty. Využívá totiž Fisherova permutačního principu. Pak hovoříme o *permutačních testech*. Tento typ testů nenabízí námi používané programy, a proto se jimi nebudeme detailněji zabývat.

Znaménkový test hodnoty mediánu má hypotézy o mediánu základního souboru, tedy

$$H_0: \tilde{\mu} = \tilde{\mu}_0 \text{ vs. } H_A: \tilde{\mu} \neq \tilde{\mu}_0.$$

Testovací statistika je založena na tom, že se spočte počet hodnot souboru větších než hodnota hypotetického mediánu Z_+ a počet hodnot menších než hodnota hypotetického mediánu Z_-. Hodnoty stejné jako medián se vynechávají. O zamítnutí nulové hypotézy se rozhoduje na základě toho, jak moc pravděpodobný je zjištěný výskyt hodnot větších než medián.

Test lze spočítat pomocí programu Statistica v nabídce Statistiky → Neparametrická statistika → Porovnání dvou závislých vzorků (proměnné). Tento nástroj není primárně určen k výpočtu testu založeného na jednom výběru, a proto požaduje proměnné dvě. Jako první proměnnou vložíme testovaný soubor a jako druhou uměle vytvořený datový soubor o stejném rozsahu jako testovaný soubor obsahující hypotetické hodnoty mediánu. Pokud např. testuji $H_0: \tilde{\mu} = 178$ a výběrový soubor má 10 hodnot, pak druhý soubor bude obsahovat 10x hodnotu 178. P-hodnotu znaménkového testu vypočteme tlačítkem Znaménkový test. Alternativně lze využít kalkulátor na adrese http://www.fon.hum.uva.nl/Service/Statistics/Sign_Test.html.

Wilcoxonův test hodnoty mediánu má hypotézy o mediánu základního souboru, tedy

$$H_0: \tilde{\mu} = \tilde{\mu}_0 \text{ vs. } H_A: \tilde{\mu} \neq \tilde{\mu}_0.$$

Testovací statistika vychází z absolutních rozdílů mezi hodnotami a hypotetickým mediánem, které seřadíme podle velikosti. Sečteme zvlášť pořadí hodnot nad hypotetickým mediánem a zvlášť pod hypotetickým mediánem. Pokud platí nulová hypotéza, měly by být oba součty pořadí přibližně stejné.

Také tento test lze spočítat v programu Statistica. Je ve stejné nabídce jako znaménkový test (Statistiky → Neparametrická statistika → Porovnání dvou závislých vzorků (proměnné)) a data je nutno připravit stejným způsobem. P-hodnotu testu vypočteme stisknutím tlačítka Wilcoxonův test.

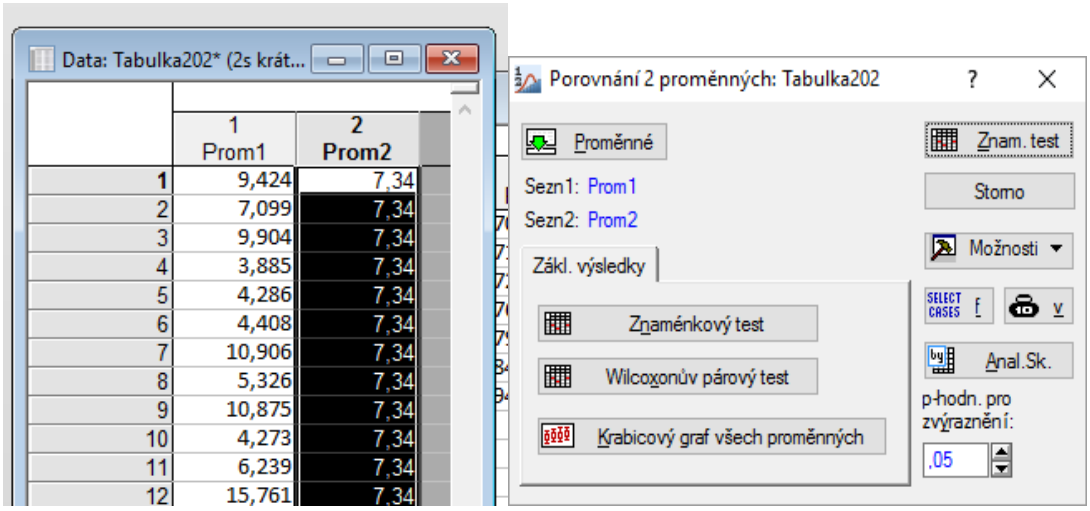
Jako další možnost testování lze využít boostrapové intervaly spolehlivosti. Jejich výhody byly již zmíněny v jedné z předchozích kapitol. Další nespornou výhodou takového postupu je skutečnost, že je lze vypočítat pro libovolný parametr rozdělení. Odpadá tedy nutnost hledání testovací statistiky pro konkrétní parametr. Při takovémto testování stanovíme požadovanou hladinu významnosti, vypočteme bootstrapový interval na stejné hladině spolehlivosti a posoudíme, zda se testovaná statistika nachází nebo nenachází ve vypočteném intervalu.

Příklad: Základní soubor dat má zešikmené chí-kvadrát rozdělení se stupni volnosti v=8. Z tohoto souboru jsme provedli náhodný výběr 20 hodnot. Medián základního souboru je 7,34, medián spočítaný z výběru je 6,62. Pomocí neparametrických testů ověříme hypotézu, že medián základního souboru je skutečně 7,34 v následujících krocích:

1. Formulace výzkumné otázky ve formě *nulové* a *alternativní hypotézy*.

$$H_0: \tilde{\mu} = 7{,}34 \text{ vs. } H_A: \tilde{\mu} \neq 7{,}34,$$

   tedy půjde o oboustranný test.
2. Volba hladiny významnosti α. Standardní hodnota při testování je $\alpha = 0{,}05$.
3. Výpočet testovacích statistik znaménkového a Wilcoxonova testu provedeme pomocí programu Statistica. Data je nutno upravit do dvou proměnných následovně:



Obr. Příprava dat a zadávací okno nástroje Porovnání dvou závislých vzorků (proměnné) programu Statistika.

4. K formulaci závěru testu využijeme p-hodnoty testů a ve druhém kroku stanovenou hladinu významnosti α. v datech. Vypočtené p-hodnoty jsou:
   Znaménkový test: p-hodnota = 0,502
   Wilcoxonův test: p-hodnota = 0,911
   Protože jsou obě p-hodnoty větší než α, nezamítáme ani v jednom případě Ho. Na hladině významnosti 0,05 tak platí, že medián základního souboru je 7,34

Hypotézu ještě posoudíme prostřednictvím bootstrapového 95% intervalu spolehlivosti. Ten pro medián základního souboru vyšel následovně:
Jednoduchý interval: <3,34; 8,77>
Kvantilový interval: <4,44; 9,94>.

Jak je vidět padne testovaná hodnota mediánu 7,34 do obou intervalů a tedy i tímto způsobem můžeme potvrdit platnost hypotézy, že medián základního souboru je 7,34.